FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

TÍTULO: Una mirada al espacio: exoplanetas K2 con imputación múltiple, PCA y ensamblado

***Alumno*: Leo Rodríguez Castilla**

***Tutor*: Fernando Pérez Contreras**

Junio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Para el Leo de 7 años:

estamos un paso más cerca de descubrir vida extraterrestre.

Resumen

La búsqueda de exoplanetas es un área fundamental en la astronomía moderna, el estudio del espacio extrasolar permite ampliar nuestro conocimiento sobre el universo. En este trabajo se presenta un análisis estadístico aplicado para la detección y clasificación de exoplanetas utilizando datos del telescopio K2 de la NASA. Se han implementado técnicas de imputación múltiple para el tratamiento de los valores perdidos, técnicas de componentes principales para manejar la multicolinealidad y se han entrenado modelos predictivos de clasificación, logística binaria, árbol de clasificación, *Random Forest* y *Extreme Gradient Boosting*, para identificar la presencia de exoplanetas en el conjunto de datos, consiguiendo capturar relaciones internas y complejas entre las variables. Los resultados muestran un rendimiento satisfactorio en términos de sensibilidad, especificidad y área bajo la curva ROC, evidenciando la capacidad de los modelos para discriminar entre objetos celestes con y sin exoplanetas. Este estudio contribuye a mejorar los métodos estadísticos en la detección de exoplanetas y abre vías para futuras investigaciones en astronomía estadística y *machine learning* aplicado.

Palabras clave: exoplanetas, misión K2 (Kepler), imputación múltiple, análisis de componentes principales, aprendizaje supervisado.

Abstract

The search for exoplanets is a fundamental field in modern astronomy, as the study of extrasolar space broadens our understanding of the universe. This paper presents a statistical analysis applied to the detection and classification of exoplanets using data from NASA's K2 telescope. Multiple imputation techniques have been implemented to address missing values, and principal component methods were applied to manage multicollinearity. Predictive classification models—including binary logistic regression, classification trees, Random Forest, and Extreme Gradient Boosting—were trained to identify the presence of exoplanets in the data set, effectively capturing internal and complex relationships among variables. The results demonstrate satisfactory performance in terms of sensitivity, specificity, and area under the ROC curve, highlighting the models' ability to discriminate between celestial objects with and without exoplanets. This study contributes to advancing statistical methods for exoplanet detection and opens pathways for future research in statistical astronomy and applied machine learning.

Keywords: exoplanets, K2 mission (Kepler), multiple imputation, principal component analysis, supervised learning.

Índice

Índice de ilustraciones

Índice de ecuaciones.

Índice de tablas y figuras.

1. Introducción

1.1 Contexto

Durante siglos se ha conjeturado sobre la existencia de planetas más allá de nuestro sistema solar, pero no fue hasta 1992 que se tuvieron indicios del primer exoplaneta (Wolszczan & Frail, 1992).

La RAE define exoplaneta como «Planeta que está fuera de nuestro sistema solar.», es decir un planeta que orbita una estrella distinta al Sol (Real Academia Española, s. f.).

La búsqueda de exoplanetas es un área de investigación activa en astronomía internacional, motivada principalmente por la búsqueda de vida extraterrestre. Tanto es así que, durante la elaboración de este trabajo se vivió un momento de euforia en el mundo de la astronomía por la posible detección de vida en el exoplaneta K2-18b (Mediavilla, 2025). Desde el descubrimiento del primero se han ido implementando nuevas técnicas y mejoras tecnológicas que permiten una mejor y más rápida detección de estos.

El primer registro que se tiene sobre la especulación de exoplanetas es del siglo XVI, siglo en el que el astrónomo Giordano Bruno conjeturó sobre que las estrellas que se vislumbraban en el cielo no eran más que otros «soles» sobre los que orbitaban otros planetas (Knox., 2024).

No fue hasta la década de 1990 en la que se hicieron los primeros descubrimientos en este ámbito, en 1992 se detectaron los primeros cuerpos astronómicos que orbitaban a una estrella pulsar[1] B1257 + 12 a aproximadamente 2000 años luz de la tierra (Wolszczan & Frail, 1992).

En 1995 Michel Mayor y Didier Queloz dieron con el primer exoplaneta, «51 Pegasi b» (Mayor & Queloz, 1995), un planeta clasificado como tipo Júpiter, desde entonces los esfuerzos por encontrar más planetas no han cesado.

Uno de los métodos más comunes para la detección de exoplanetas se denomina método de tránsito (Ortega, 2021), consiste en observar las variaciones de luz emitidas por una estrella debido al tránsito de un planeta, variaciones que son muy sutiles y están enmascaradas por diferentes ruidos entre los que se encuentran el ruido instrumental, el ruido estelar y ruido atmosférico en las observaciones terrestres. La estadística permite 1) trabajar con una gran cantidad de datos y 2) discernir entre tránsitos planetarios y ruidos espurios. Además, tras la

[1] Tipo de estrella que emite radiación periódica debido a su rápida rotación y fuerte campo magnético

detección de un exoplaneta se realizan estimaciones de sus parámetros físicos, lo que genera una incertidumbre que la estadística es capaz de medir.

Este estudio se enfoca por un lado en entender las características comunes de los parámetros estelares de los exoplanetas detectados por el telescopio K2 y, por otro lado, se han realizado modelos de predicción con el objetivo de entender que parámetros podrían indicar la existencia de exoplanetas en otros sistemas.

En el mes de mayo de 2025, el NASA Exoplanet Archive tras varias misiones estelares, entre las que se encuentran la misión Kepler y su continuidad K2, confirmó un total de 5889 exoplanetas de diversas características. Según el Exoplanet Archive la mayoría de los planetas extrasolares tienen un radio entre 1.25 y 6 radios terrestres ($R_\oplus$)

Rango de radio ($R_\oplus$)	Número de exoplanetas
$R \leq 1.25R_\oplus$	541
$1.25R_\oplus < R \leq 2R_\oplus$	1093
$2R_\oplus < R \leq 6R_\oplus$	1889
$6R_\oplus < R \leq 15R_\oplus$	675
$15R_\oplus < R$	213

Tabla 1: distribución del radio de los exoplanetas

Nota. *Datos extraídos del NASA Exoplanet Archive (2025). Disponible en https://exoplanetarchive.ipac.caltech.edu*

1.2 Objetivos

El objetivo general de este trabajo se centra en construir y evaluar un clasificador que estime la probabilidad de que un objeto observado por el telescopio K2 sea un exoplaneta mediante propiedades estelares y fotométricas.

En cuanto a objetivos específicos se han desarrollado los siguientes:

1. Comparar modelos clásicos con modelos más complejos mediante validación cruzada repetida estratificada.
2. Evaluar el poder predictivo de las variables fotométricas y parámetros estelares para discriminar entre exoplaneta y no exoplaneta.
3. Aplicar imputación múltiple y combinar estimaciones mediante las reglas de Rubin para estimar la variabilidad entre imputaciones.

1.3 Metodología

El trabajo se ha realizado íntegramente en el software estadístico Rstudio.

Primero se realizó un análisis descriptivo exploratorio de las variables, lo que sirvió para entender la relación de las variables con la variable dependiente, a continuación, se depuró la base de datos, eliminando observaciones duplicadas o erróneas, tratando los valores atípicos (*outliers*) e imputando los valores perdidos mediante imputación múltiple.

La variable dependiente del estudio, dummy_disposition, fue creada a partir de la variable *disposition*, de naturaleza categórica, se clasificaron las categorías de manera dicotómica transformando el objetivo del estudio a un problema de clasificación binaria.

Para el estudio de la relación latente de los datos, así como para la obtención de perfiles de interés se entrenó un modelo de regresión logística binaria y un árbol de clasificación. Con intención de mejorar la capacidad predictiva de los modelos se optó por modelos más complejos como *Random Forest* y XGBoost.

La validación de todos los modelos se hizo mediante validación cruzada repetida y estratificada por sistema planetario, de esta forma se pudieron conseguir métricas fiables y comparables entre modelos.

2. Marco teórico

Como primer enfoque se desarrolló un modelo de logística binaria, con el objetivo de poder entender la función latente entre los datos y obtener *odds ratio* y coeficientes interpretables físicamente.

El modelo de logística binaria sirve para entender cómo funcionan las relaciones en un conjunto de variables independientes sobre una variable cualitativa dicotómica, es decir, que toma dos posibles valores. Es una generalización del modelo lineal que usa la función logística como enlace (Ayçaguer & Utra, 2004, págs. 11-13, 53 -57).

Este modelo permite calcular probabilidades de ocurrencia sobre la variable dependiente, tomando los valores Y=1 o Y = 0, es decir, ocurrencia o no ocurrencia del evento. Se predecirá que ocurrirá el evento si se supera el umbral (por defecto, 0.5), en caso contrario se predecirá que no ocurrirá el evento. Si la variable dependiente tiene proporciones evento y no evento muy desbalanceadas, se puede cambiar el umbral para aumentar la capacidad de discriminación del modelo.

La probabilidad de que ocurra el evento viene dada por la Ecuación 1:

$$p = P(Y = 1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 - \cdots \beta_j X_j}} = \frac{1}{1 + e^{-Z}}$$

Ecuación 1: Probabilidad de que ocurra el evento en el modelo logístico

El modelo logístico es interesante, porque, permite calcular los *odds ratio* que vienen dados por la razón entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra, como se muestra en la ecuación 2.

$$Odds(evento) = \frac{P(evento)}{1 - P(evento)}$$

Ecuación 2: Cálculo de odds ratio

Para una variable cuantitativa si el *odds ratio* toma valores superiores a uno, significa que la probabilidad de que ocurra el evento aumenta en función de las unidades de esta variable, por otro lado, si toma valores inferiores a uno, significa que la probabilidad disminuye en función de las unidades de la variable. En el caso de que la variable predictora sea cualitativa, el *odds ratio* indicará un aumento o descenso de la probabilidad de éxito en caso de cambio entre categorías.

Para añadir calidad interpretativa al estudio, así como perfiles de interés, se entrenó un árbol de clasificación, este modelo es interesante por su facilidad interpretativa, así como su capacidad de captar relaciones no lineales entre las variables.

Los árboles de decisión están formados por las siguientes partes: nodo principal o raíz el cual contiene la totalidad de la muestra, nodos intermedios que generan las subdivisiones de la muestra según las variables explicativas y por último los nodos finales u hojas.

Conviene distinguir entre los dos tipos de árboles de decisión: regresión y clasificación. El árbol será de regresión cuando la variable dependiente sea de naturaleza cuantitativa, por otro lado, el árbol será de clasificación cuando la variable de estudio sea de naturaleza cualitativa (López, 2011, pág. 162).

El árbol de clasificación se creó mediante el algoritmo *Classification And Regression Trees*, en adelante CART, (implementado en el paquete rpart de R), CART hace divisiones binarias. Como criterio de división se contrastó entre el índice de Gini y la entropía. En cada nodo t se selecciona la variable y el punto de corte que maximiza la reducción de impureza. La impureza de Gini viene definida en la Ecuación 3, donde $p_k(t)$, es la proporción de la clase k en el nodo t.

$$G(t) = 1 - \sum_{k=1}^{K} p_k(t)^2$$

Ecuación 3: cálculo del índice de Gini.

La impureza según la entropía viene definida en la ecuación 4.

$$H(t) = -\sum_{k=1}^{K} p_k(t) log_2 p_k(t)$$

Ecuación 4: cálculo de la entropía

Es importante destacar que ambos criterios suelen dar divisiones similares, aunque no idénticas.

Dado un candidato de división s, la calidad se mide por la ecuación 5, que se puede resumir como la impureza del padre menos la impureza media ponderada de los hijos. Donde $I(t)$ es la impureza de Gini o entropía en el nodo actual t, t_L y t_R son los nodos hijos y n_L y n_R es el número de observaciones en los nodos hijos.

$$\Delta I(s,t) = I(t) - \frac{n_L}{n_t} I(t_L) - \frac{n_R}{n_t} I(t_R)$$

Ecuación 5: Expresión de la reducción de impureza

Para ajustar la complejidad del árbol se tuneó el *complexity parameter* (cp), con valores que van desde 0.001 hasta 0.02, y se eligió aquel valor que maximizó el área bajo la curva de *Receiver Operating Characteristic* (ROC) en validación cruzada; Para evitar hojas con casos mínimos se estableció el parámetro *minbucket* (que controla el número mínimo de observaciones que debe tener una hoja final) en 20.

Los árboles de decisión a pesar de ser modelos muy accesibles tienen como inconveniente una alta variabilidad, un cambio en los datos de entrada puede generar un árbol muy distinto, esto puede derivar en desconfianza sobre sus predicciones. Por ello, Breiman, a finales del siglo XX y principios del siglo XXI presentó los modelos de ensamblado: *Bagging* (Breiman, 1996) y posteriormente *Random Forest* (Breiman, 2001). Estos modelos promedian una gran cantidad de árboles para reducir la varianza y mejorar la generalización de las predicciones. En la misma línea el *gradient boosting* (Friedman, 2001) construye árboles secuencialmente, corrigiendo los errores cometidos por los árboles anteriores, años después apareció XGBoost, una

implementación optimizada del mismo algoritmo (Chen & Guestrin, 2016) que añade regularización y optimización computacional.

Los modelos de ensamblado combinan las predicciones de muchos modelos base (más sencillos), obteniendo así, una predicción final más estable. Existen dos categorías fundamentales: (1) reducción de varianza, modelos como *bagging* y *Random Forest* y (2) reducción de sesgo, modelos de mejora secuencial como *CatBoost* o *gradient boosting*

El modelo *Random Forest* se creó como ensamblado de árboles CART entrenados con muestreo *bootstrap* y selección aleatoria de variables. Para conseguir una buena calidad predictiva se tunearon varios parámetros:

1. *Mtry*, es decir, el número de variables que se pueden probar en cada partición. Se probaron tres posibles valores: $\frac{\sqrt{(p)}}{2}$, $\sqrt{(p)}$, $2\sqrt{(p)}$
2. *Splitrule*, es decir, la regla de división que se probó entre Gini y *extratrees*.
3. *Min.node.size*, es decir, el tamaño mínimo que debe tener un nodo para poder dividirse. Se probaron cinco posibles valores: 1, 5, 10, 20 y 30.

Además, como las clases de la variable dependiente estaban desbalanceadas se realizó un balance de clases. A la clase minoritaria se le asignó un peso inverso a su prevalencia. De esta forma el modelo no «ignora» la clase minoritaria.

El modelo XGBoost, como se nombra anteriormente, es una implementación optimizada de *gradient boosting*. Como el resto de los métodos de *boosting* construye de manera secuencial muchos árboles sencillos o *weak learners* que van corrigiendo los errores de los anteriores buscando optimizar una función de pérdida, en el caso de la clasificación la función logística.

Con intención de optimizar el tiempo de cómputo, la búsqueda del mejor modelo mediante el tuneo de hiperparámetros se realizó en dos fases.

Una primera en la que se tunearon los siguientes hiperparámetros:

1. *Nrouds*, es decir, el número de iteraciones. Se probaron dos valores: 200 y 400
2. *Max_depth,* es decir, la complejidad de los árboles. Se probaron dos valores 3 y 4
3. *Learning rate (η).* Se fijó el valor en 0.1
4. *Gamma,* es decir, la reducción mínima de pérdida necesaria para permitir un *split*, se fijó en 0
5. *Colsample_bytree,* es decir, la proporción de variables muestreadas se fijó en 0.8

6. *Min_child_weight*, es decir, la cantidad de «masa» efectiva de un nodo hijo. Se probó con 1 y 3
7. *Subsample,* es decir, el tamaño de la submuestra seleccionada. Se fijó en 0.8

La combinación seleccionada fue la que maximizó el área bajo la curva ROC.

Más tarde en la segunda etapa se añadieron unos valores alrededor de la combinación ganadora de la primera etapa. Como se expondrá más adelante.

Debido al tamaño reducido de la muestra se optó por no dividir la base de datos en prueba y entrenamiento, si no que la evaluación de todos los modelos se hizo mediante validación cruzada k-fold.

Para evaluar los modelos, primero se graficó la curva ROC que para todos los umbrales de decisión enfrenta sensibilidad contra 1 – especificidad.

La sensibilidad es la probabilidad de clasificar como éxito a una observación que realmente es exitosa, la fórmula se presenta en la ecuación 6, donde VP es verdadero positivo y FN falso negativo.

$$sensibilidad = \frac{VP}{VP + FN}$$

Ecuación 6: cálculo de sensibilidad

Por otro lado, la especificidad se define como la probabilidad de clasificar como no éxito a una observación que realmente no es exitosa, la fórmula se presenta en la ecuación 7, donde VN es verdadero negativo y FP es falso positivo.

$$especificidad = \frac{VN}{VN + FP}$$

Ecuación 7: cálculo de especificidad

A la sensibilidad y especificidad se le acompañará del valor predictivo positivo (PPV), que se define como la probabilidad de que un objeto predicho como exoplaneta realmente lo sea y valor predictivo negativo (NPV), que se define como la probabilidad de que un objeto definido como no exoplaneta realmente lo sea.

$$PPV = \frac{VP}{VP + FP}$$

Ecuación 8: cálculo del valor predictivo positivo

$$NPV = \frac{VN}{VN + FN}$$

Ecuación 9: cálculo del valor predictivo negativo

Donde VP es verdadero positivo, FP es falso positivo, VN es verdadero negativo y FN falso negativo.

Estos cálculos se pueden realizar con un punto de corte 0.5, pero si la variable dependiente tiene clases muy desbalanceadas, sería posible cambiar el punto de corte para mejorar la calidad predictiva.

Tras graficar la curva ROC, se calcula el área bajo la curva (AUC) que toma valores entre 0.5 y 1; si el modelo predice como el azar el AUC tomará un valor en torno de 0.5, si el modelo discrimina perfectamente, el AUC rondará el valor 1.

En el trabajo se presentarán tanto las métricas con el punto de corte 0.5 así como con el punto óptimo de Youden, el cuál maximiza sensibilidad y especificidad.

El cálculo para hallar el punto óptimo de Youden se presenta en la ecuación 10, este cálculo se ha de hacer para cada posible umbral y se escogerá aquel que maximice Y_i

$$Y_i = sensibilidad + especificidad - 1$$

Ecuación 10: cálculo de índice de Youden

3. Datos y preparación

3.1 Descripción de la base de datos

La detección de exoplanetas es un área relativamente reciente tanto en el mundo de la física como en el de la estadística. Hasta la fecha de esta consulta – 28 de agosto de 2025- se ha confirmado la existencia de menos de 6000 exoplanetas (NASA Exoplanet Science Institute, 2025). La NASA ha enviado al espacio tres grandes misiones dedicadas al estudio de los planetas extrasolares: Kepler con 2784 exoplanetas confirmados, K2 con 547 exoplanetas confirmados y TESS con 686.

Para el estudio se han escogido los datos de tabulados de la misión estelar K2, misión que dio continuidad y mejoró el trabajo del telescopio Kepler, entre otras cosas, observó una mayor parte del cielo, lo que dio la oportunidad de capturar distintos entornos estelares. El telescopio

K2 ofrece una muestra más heterogénea que su predecesor, reduciendo de esta forma el sesgo del estudio (Howell, y otros, 2014), (Mightell & Van Cleve, 2020).

El conjunto de datos escogido contiene mediciones estelares y fotométricas con identificadores de objetos observados por el telescopio K2.

Toda la información utilizada en el análisis corresponde a la extracción realizada el 28 de marzo de 2025. Las unidades y definiciones de cada variable seleccionada para el estudio se detallan en el apartado 3.2.

A partir de la base de datos original obtenida del Exoplanet Archive[2], que incluye 3893 registros, se eliminaron aquellas observaciones no etiquetadas como representativas, mediante el uso de la variable auxiliar deafult_flag, que toma el valor 1 para aquellas observaciones calificadas como relevantes, de esta manera se obtuvo un total de 1803 observaciones únicas. A continuación, se retiraron del estudio los objetos clasificados como candidatos, y, finalmente se eliminaron las observaciones con datos perdidos en la variable dependiente (dummy_disposition) o en todas las variables, quedando finalmente una muestra depurada de 827 objetos estelares únicos. Estos objetos están clasificados según su naturaleza en exoplanetas o no exoplanetas. Cabe destacar que esta clasificación y selección se realizó específicamente para los fines de este análisis.

3.2 Variables seleccionadas y definición

El conjunto final incluye 26 variables agrupadas en (i) fotométricas, (ii) estelares y (iii) auxiliares. La variable dependiente del estudio, llamada *dummy_disposition*, fue creada a partir de la variable *disposition*, que inicialmente contaba con 4 categorías: *Candidate, False Positive, Confirmed* y *Refuted*. Como se menciona anteriormente, los objetos estelares clasificados como candidatos quedaron fuera del análisis, de las categorías restantes se definió

- Exoplaneta como *Confirmed*
- No exoplaneta como la suma de observaciones de *False Positive* y *Refuted.*

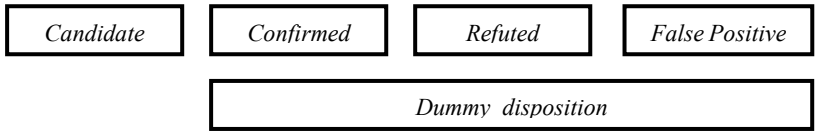


Figura 1: Proceso de creación de la variable dependiente

[2] Conjunto de datos disponible en: https://exoplanetarchive.ipac.caltech.edu

A continuación, se presenta la explicación de las variables que forman parte del estudio

- pl_name: identificador único para cada objeto estelar.
- st_teff: temperatura de la estrella modelada como un cuerpo negro, en grados Kelvin.
- st_rad: distancia desde el centro de la estrella hasta su superficie, medido en radios solares.
- st_mass: cantidad de materia contenida en la estrella, en masas solares.
- st_met: medida del contenido metálico de la estrella en comparación con el contenido de hidrógeno, medida en unidades logarítmicas (dex). Un valor positivo indica que la estrella es más metálica que el Sol y un valor negativo, lo contrario.
- st_logg: mide la aceleración gravitatoria que se experimenta en la superficie de la estrella, medida en logaritmo de la gravedad en cm/s^2.
- sy_pm: velocidad angular aparente con la que estrella o el sistema planetario se desplaza a través del cielo, observado desde el centro de masa del Sistema Solar. Medido en milisegundos de arco por año.
- sy_dist: distancia al sistema planetario en parsecs[3].
- sy_bmag: brillo aparente del sistema planetario medido en la banda B del sistema fotométrico Johnson, que corresponde a luz azul visible (~440 nm).
- sy_vmag: brillo aparente en la banda V del sistema Johnson, centrada en luz visible verde (~550 nm). Es una de las magnitudes estándar más usadas.
- sy_jmag: brillo aparente medido en la banda J del infrarrojo cercano (~1.25 μm) según el catálogo 2MASS.
- sy_hmag: brillo en la banda H del infrarrojo cercano (~1.65 μm), también del catálogo 2MASS.
- sy_kmag: brillo en la banda Ks (~2.16 μm), parte del infrarrojo cercano, utilizada para estudiar estrellas y objetos fríos.
- sy_umag: brillo medido en la banda ultravioleta u del sistema Sloan Digital Sky Survey (SDSS), centrada alrededor de 355 nm.
- sy_gmag: magnitud en la banda g (~475 nm), que corresponde a luz visible azul-verde.
- sy_rmag: brillo en la banda r (~622 nm), luz visible roja.
- sy_imag: magnitud en banda i (~763 nm), zona del espectro cercana al infrarrojo.
- sy_zmag: magnitud en banda z (~913 nm), más cercana al infrarrojo que la i.
- sy_w1mag: brillo en la banda W1 del telescopio espacial WISE, centrada en el infrarrojo medio.
- sy_w2mag magnitud en la banda W2 de WISE, también en el infrarrojo medio.
- sy_w3mag: brillo en la banda W3 de WISE, infrarrojo térmico.
- sy_w4mag magnitud en la banda W4, la más larga de WISE, en el infrarrojo lejano.
- sy_gaiamag: brillo aparente medido por el satélite Gaia en su banda amplia que cubre del visible al infrarrojo cercano.
- sy_tmag: magnitud aparente medida en la banda del telescopio espacial TESS, diseñado para detectar exoplanetas.
- sy_kepmag: brillo aparente medido en la banda del telescopio Kepler, optimizada para detectar tránsitos planetarios.

[3] Un parsec es una unidad de distancia astronómica equivalente a 3.26 años luz, se define como la distancia a la cual una estrella muestra una paralaje anual de un segundo de arco

- dummy_disposition: estado de clasificación del objeto, 1 para exoplanetas y 0 para no exoplanetas

3.3 Análisis descriptivo exploratorio EDA

3.3.1 Análisis univariante

En esta sección se muestra un resumen estadístico de las variables continuas del estudio con el fin de entender la distribución de cada variable y detectar posibles anomalías. Para cada variable se calcularon la media, mediana, la desviación típica, el rango intercuartílico, el máximo y el mínimo, la Tabla 2 muestra los estadísticos descriptivos correspondientes.

variable	media	mediana	sd	IQR	min	max	asimetria
st_teff	5165.69	5327.16	1031.61	1275.75	2566	11886	0.35
st_rad	1.08	0.91	0.73	0.55	0.12	7.82	3.48
st_mass	0.88	0.88	0.32	0.32	0.09	3.16	1.04
st_met	-0.06	-0.03	0.22	0.26	-1.2	0.53	-0.8
st_logg	4.43	4.48	0.31	0.37	3	5.24	-0.75
sy_pm	53.24	26.54	110.65	36.95	0.4	1026.58	6.85
sy_dist	428.37	273.8	501.87	330.9	21.82	5200	3.67
sy_bmag	14.12	13.8	2.14	2.65	5.84	20.36	0.11
sy_vmag	13.27	13.02	1.97	2.42	5.84	20.25	0.07
sy_jmag	11.41	11.41	1.53	1.85	5.76	16.67	-0.17
sy_hmag	10.97	10.98	1.49	1.76	5.64	16.41	-0.11
sy_kmag	10.85	10.88	1.49	1.73	5.57	15.74	-0.12
sy_umag	16.58	15.86	2.02	2.14	14.03	30	1.96
sy_gmag	14.72	14.82	2.03	2.63	11.16	30	1.46
sy_rmag	13.83	13.63	1.89	2.32	9.9	27.42	1.43
sy_imag	13.5	13.4	1.7	2.3	10.33	22.54	0.89
sy_zmag	13.49	13.43	1.22	0.97	9.85	21.8	1.6
sy_w1mag	10.81	10.81	1.46	1.71	5.58	16.19	-0.12
sy_w2mag	10.83	10.85	1.45	1.69	5.42	15.94	-0.14
sy_w3mag	10.63	10.77	1.23	1.64	5.62	12.9	-0.83
sy_w4mag	8.59	8.67	0.42	0.47	5.54	9.34	-2.34
sy_gaiamag	12.96	12.81	1.83	2.24	5.81	20.22	0.04
sy_tmag	12.32	12.25	1.69	2.13	5.82	20.4	0

Tabla 2: estadísticos descriptivos de variables continuas

Como vemos en la Tabla 2, las variables sy_pm y sy_dist presentan medias muy superiores a sus medianas, indicando una distribución claramente sesgada hacia valores altos. En consecuencia, se calculó el coeficiente de asimetría, mostrado en la ecuación 11.

$$\gamma = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{\frac{3}{2}}}$$

Ecuación 11: cálculo de la asimetría

Se obtuvieron valores mayores a 3 para las variables st_rad, sy_pm y sy_dist, por tanto, sus distribuciones serán tratadas en escala logarítmica. También se puede apreciar que en casi la totalidad de las bandas fotométricas (variables con terminación mag) el rango intercuartílico varía entre una y dos unidades, es decir, los valores se concentran en torno a la media.

Se observa también una mayor presencia de variables leptocúrticas, es decir, variables con colas extensas (p. ej., st_rad ≈ 20.85, sy_dist ≈ 19.63) lo que índicas colas pesadas y, por tanto, posible heterogeneidad (estrellas enanas con gigantes gaseosos). Esto podría suponer un problema en la modelización por la presencia de *outliers*, por ello, se realizará una estandarización robusta por el método de la desviación absoluta mediana (MAD) más adelante.

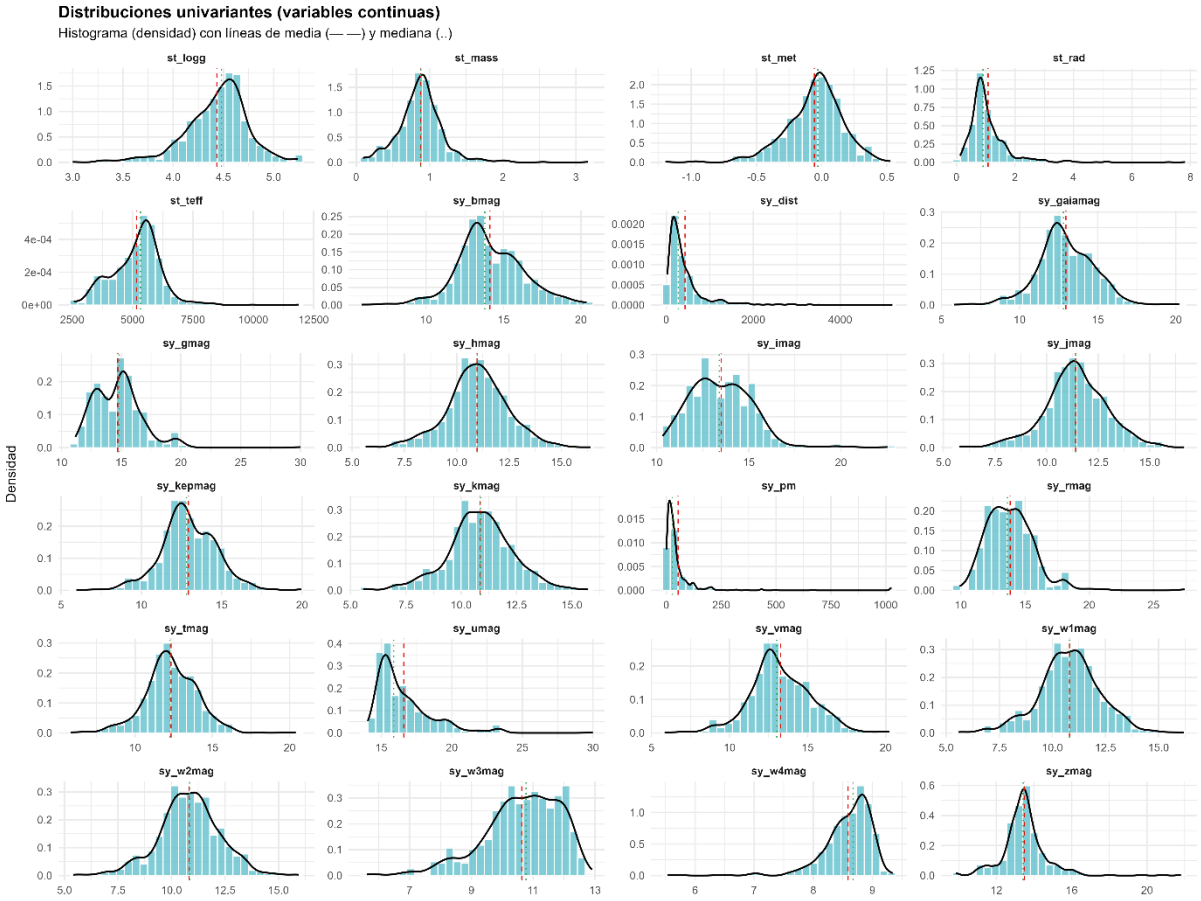


Ilustración 1: distribución de variables continuas

Se pueden ver valores fuera del rango operativo en algunas variables, como valores superiores a 25 en alguna banda fotométrica. Estos valores fueron transformados a valores perdidos para su posterior imputación.

A continuación, se estudiará la radiografía de la variable dependiente del estudio.

variable	n_niveles	nivel_mas_frecuente	frecuencia
dummy_disposition	2	Exoplaneta	577

Tabla 3: estadísticos descriptivos de la variable dependiente

Como se observa tanto en la Tabla 3 como en la ilustración 2; la variable dependiente, dummy_disposition, está claramente desbalanceada hacia la categoría «exoplaneta» con aproximadamente el 70% de los casos, esto no refleja la proporción real de exoplanetas en el espacio, si no que para la muestra se almacenan solo los datos de los objetos catalogados como exoplanetas, falsos positivos o refutados. El desbalance de la variable dependiente se tendrá en cuenta durante el proceso de entrenamiento los modelos.

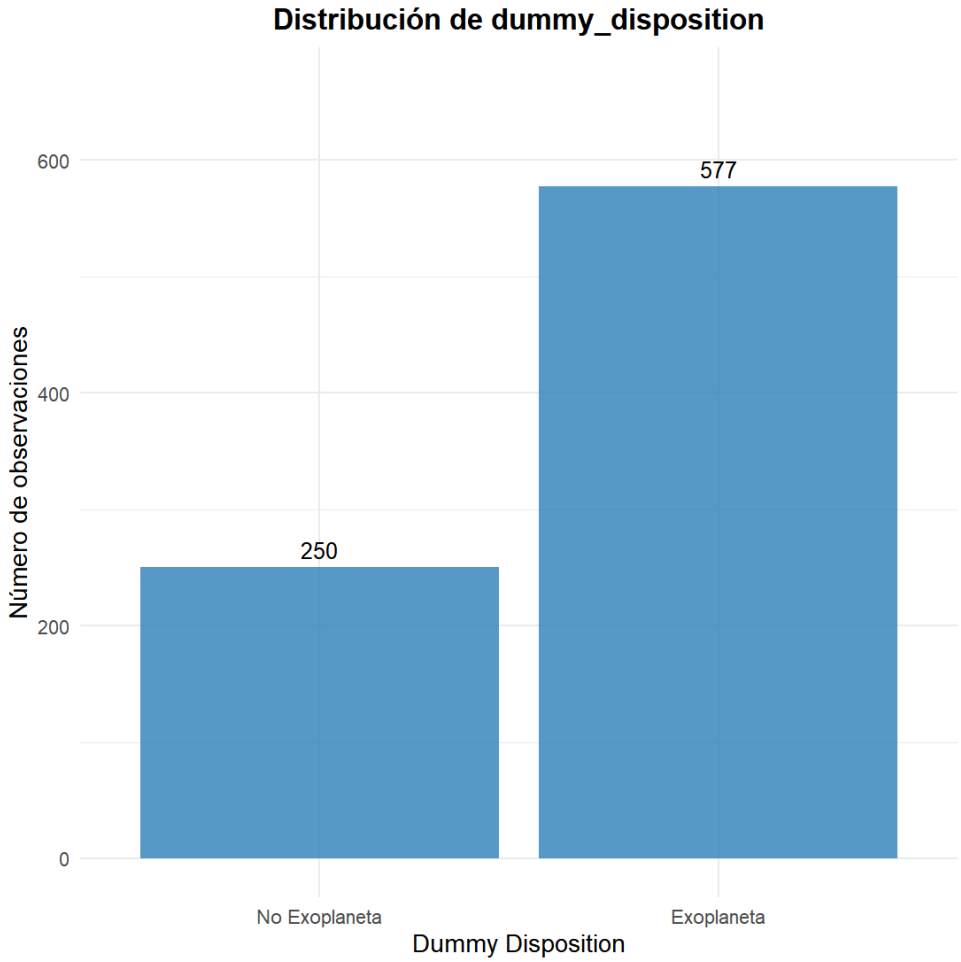


Ilustración 2: distribución de la variable dependiente

3.3.2 Análisis Bivariante

En esta sección se realiza un análisis bivariado centrado en la relación de las variables independientes con la variable dependiente del estudio: dummy_disposition (Exoplaneta vs No Exoplaneta).

El objetivo del análisis bivariado es detectar patrones y relaciones entre estas variables que permitan formular hipótesis sobre qué factores están relacionados con la presencia (o ausencia) de exoplanetas. Los contrastes t de Welch (H_0:Igualdad de medias) realizados en esta sección se realizaron con casos completos por variable. El tamaño muestral de cada grupo oscila entre 100 y 247 para la categoría «No Exoplaneta» y entre 418 y 523 para la categoría «Exoplaneta».

En la ilustración 3 se compara la velocidad angular aparente de la estrella anfitriona según el estado de confirmación, se observa una diferencia en las medias entre ambos grupos

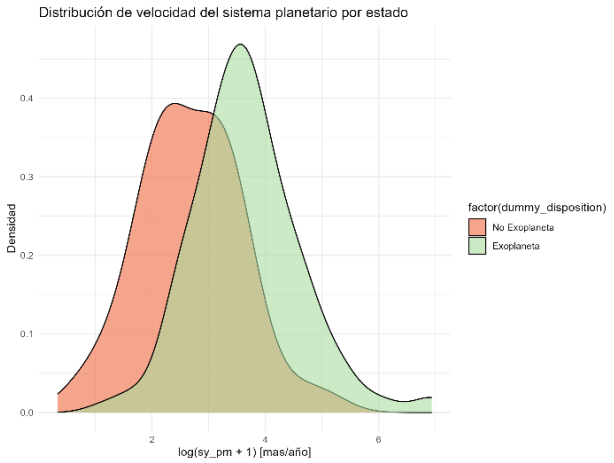


Ilustración 3: distribución de velocidad angular de la estrella según estado de confirmación

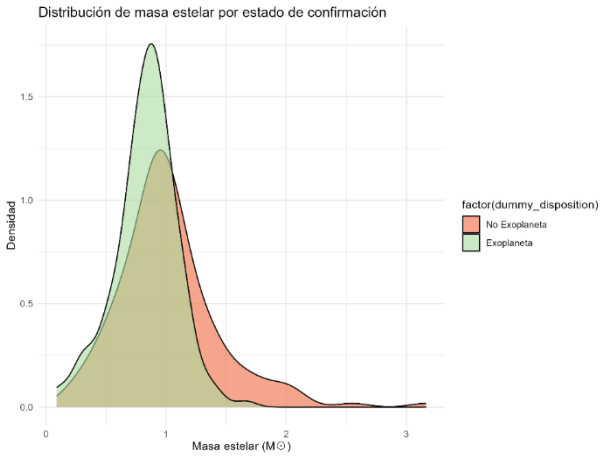


Ilustración 4 : distribución de masa estelar según el estado de confirmación

En la ilustración 4 se muestra la distribución de las masas estelares en ambos grupos de la variable dependiente (exoplaneta y no exoplaneta), parece que a mayor masa tenga la estrella anfitriona menor presencia de exoplanetas existe en el sistema.

Ambas hipótesis se contrastaron mediante la prueba t de Welch. Para la variable sy_pm, se obtuvo un estadístico $t = -14.08$, un intervalo de confianza $(-1.128, -0.852)$ y un p-valor = $2.2 \cdot 10^{-16}$. En el caso de la variable st_mass, se obtuvo un estadístico t = 5.34, un intervalo de confianza (0.118, 0.256) y un p-valor $= 2.39 \cdot 10^{-7}$. Significación suficiente en ambos casos para rechazar la hipótesis nula de igualdad de medias, por tanto, se puede concluir que a mayor velocidad angular aparente se observe en la estrella anfitriona, mayor probabilidad de pertenecer a la clase Exoplaneta. Por otro lado, a una mayor masa de la estrella anfitriona, se observará una menor probabilidad de pertenecer a la clase Exoplaneta.

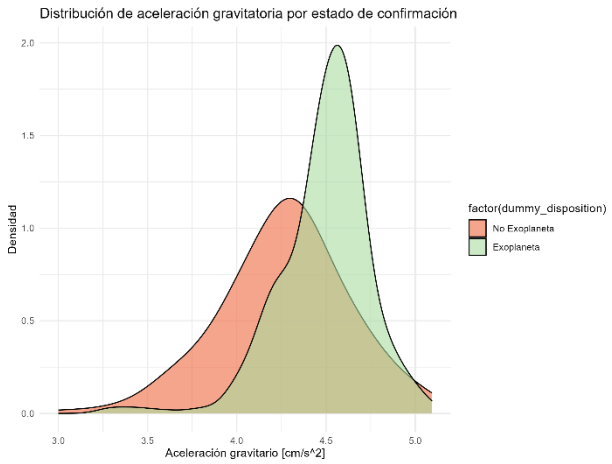


Ilustración 6: distribución de la aceleración gravitatoria según el estado de confirmación

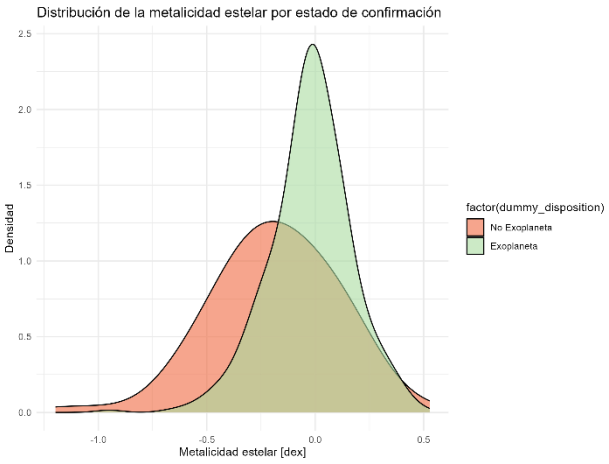


Ilustración 5: distribución de la metalicidad estelar según el estado de confirmación

Se continuó estudiando las variables st_logg y st_met. En la ilustración 5 y en la ilustración 6 se presenta la distribución de la metalicidad estelar y de la aceleración gravitacional de la estrella anfitriona respectivamente.

Para el contraste realizado en la variable st_logg, se obtuvo un estadístico t = -7.61, un intervalo de confianza al 95% de $(-0.283, -0.166)$ y un p-valor $= 4.59 \cdot 10^{-13}$, lo que demuestra que sí existen diferencias en las medias de la aceleración gravitacional según el estado de confirmación, aunque no son muy notables. Obtenemos conclusiones muy parecidas en el caso de la metalicidad estelar, ya que el contraste de la t de Welch devolvió un estadístico t = -5.4, un intervalo de confianza de $(-0.219, -0.101)$ y un p-valor $= 3.28 \cdot 10^{-7}$.

Para las variables st_rad y st_teff también se observaron las distribuciones según los distintos estados de confirmación

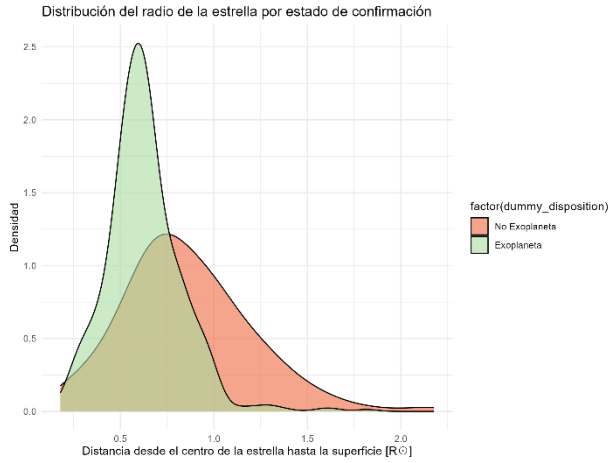


Ilustración 7: distribución del radio de la estrella según el estado de confirmación

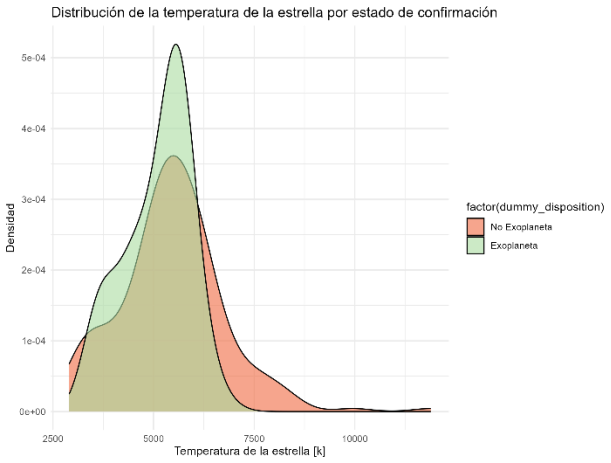


Ilustración 8: distribución de la temperatura estelar según el estado de confirmación

Como se puede observar en la ilustración 7, parece notarse que a menor radio estelar hay una mayor presencia de exoplanetas. Esta hipótesis se contrastó con la prueba de t de Welch, obteniéndose un p-valor $= 2.2 \cdot 10^{-16}$, un estadístico t = 9.14 y un intervalo de confianza al 95% de (0.178, 0.276). Evidencia suficiente para rechazar la hipótesis nula de igualdad de medias y concluir que, a mayor radio estelar, se espera encontrar una menor presencia de exoplanetas

En el caso de la variable st_teff, se puede comprobar en la ilustración 8 que gráficamente no se logra ver una gran diferencia de temperatura media según el estado de confirmación.

Se realizó la prueba de t de Welch y se obtuvo un estadístico t = 3.75, un intervalo de confianza de (175.011, 561.686) y un p-valor $= 0.0002$, evidencia suficiente para afirmar que sí existe una diferencia de medias en la temperatura estelar según el estado de confirmación, a mayor temperatura tenga la estrella anfitriona se espera encontrar una menor presencia de exoplanetas.

La última variable relacionada con los sistemas estelares que queda por estudiar es sy_dist, en la ilustración 9 se puede apreciar una clara diferencia entre las log-distancias al sistema planetario según el estado de confirmación.

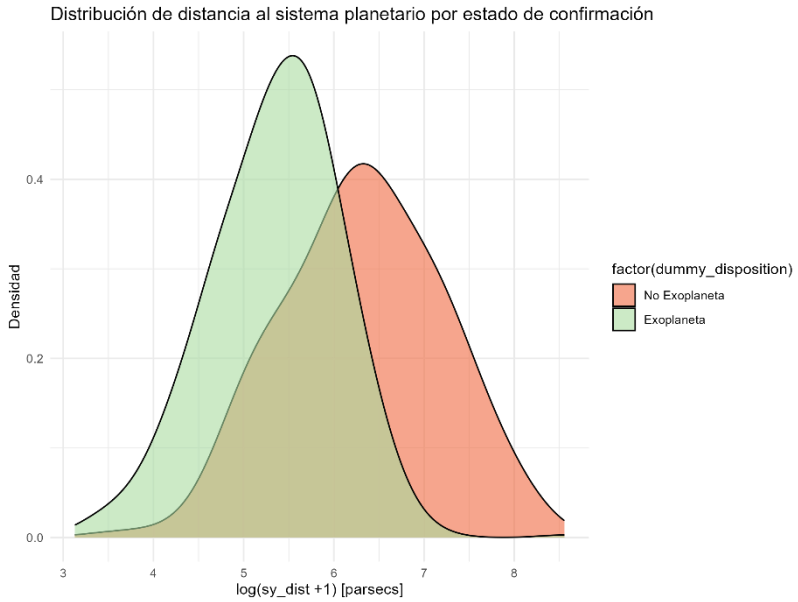


Ilustración 9: distribución de la log-distancia al sistema planetario según el estado de confirmación.

Se realizó la prueba t de Welch y se obtuvo un estadístico t = 14.94, un intervalo de confianza de (0.839, 1.094) y un p-valor = $2.2 \cdot 10^{-16}$. Por tanto, para cualquier nivel de significación se puede rechazar la hipótesis nula de igualdad de medias, concluyendo que a mayor log-distancia

al sistema planetario, se espera encontrar una menor probabilidad de pertenecer a la clase Exoplaneta.

Una vez se estudiaron las variables relacionadas con las estrellas anfitrionas y los sistemas se procedió al estudio de las bandas fotométricas. A continuación, se presentarán los cuatro gráficos con mayor valor para el trabajo, el resto pueden consultarse en el anexo.

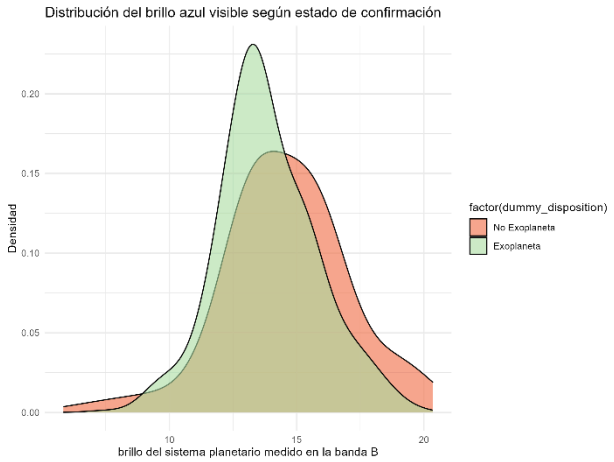


Ilustración 10: distribución de la radiación en la banda B según el estado de confirmación

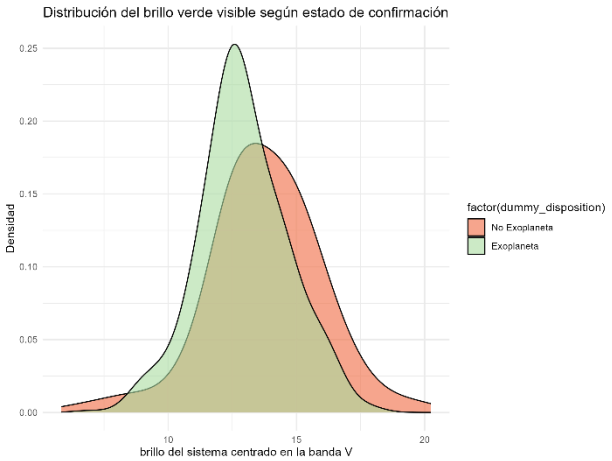


Ilustración 11: distribución de la radiación en la banda V según el estado de confirmación

En la ilustración 10 y en la ilustración 11, se observa la distribución de la banda B y V según el estado de confirmación. No se aprecian grandes diferencias en las medias gráficamente, pero en el contraste t de Welch para la igualdad de medias se obtuvieron estadísticos t = 4.97 y t = 5,54 respectivamente con ambos p-valores muy próximos a cero, por tanto, hay evidencias suficientes para rechazar la igualdad de medias y concluir que existen diferencias en el estado de confirmación según la radiación en la banda B y V.

Para la banda B se obtuvo un intervalo de confianza al 95% de (0.545, 1.26) y para la banda V se obtuvo un intervalo de confianza al 95% de (0.587, 1.233), es decir, que a mayor magnitud (menos brillo) tanto en la banda B como en la V se espera una menor presencia de exoplanetas.

A continuación, se presenta la distribución de la variable sy_tmag y sy_gaiamag según la clasificación de dummy_dispostion.

En la ilustración 12 en la que se presenta la distribución de la banda TESS no se aprecia una clara diferencia entre las distribuciones de ambos grupos (exoplaneta y no exoplaneta). Conclusiones similares se pueden sacar de la ilustración 13 en la que se expone la banda Gaia.

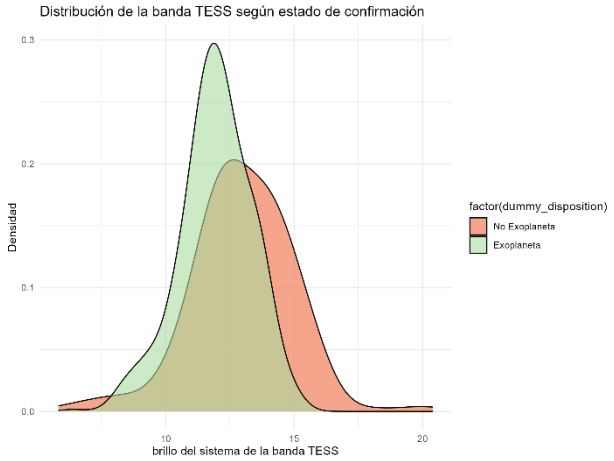


Ilustración 12: distribución de la magnitud TESS según el estado de confirmación

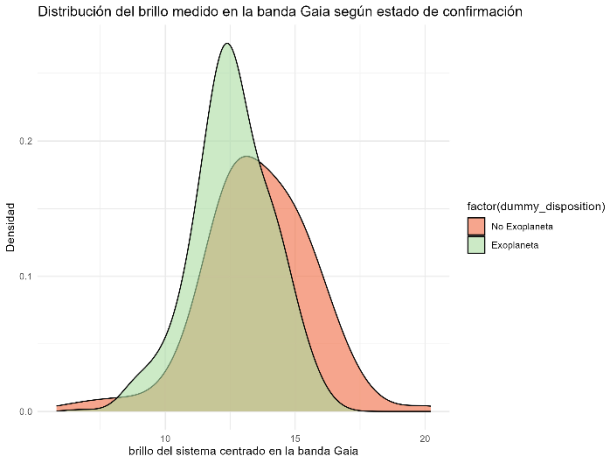


Ilustración 13: distribución de la banda Gaia según el estado de confirmación

Aunque gráficamente no se hayan observado diferencias significativas, el contraste t de Welch reportó un t = 8.057 y un p-valor de $1.096 \cdot 10^{-14}$ para la banda TESS y un estadístico t = 7.049 y un p-valor de $9.14 \cdot 10^{-12}$. Ambos significativos para cualquier nivel de significación, por ello, se concluye que existen evidencias suficientes en contra de la hipótesis nula de igualdad de medias y podemos concluir que existen diferencias la distribución de las bandas TESS y Gaia según el estado de confirmación.

Se obtuvo un intervalo de confianza al 95% para la banda TESS: (0.852, 1.403), por tanto, a mayor magnitud (menor brillo) en la banda especializada TESS, se espera encontrar una menor presencia de exoplanetas

En el caso de la banda Gaia, se obtuvieron resultados similares, con un intervalo de confianza de (0.78, 1.383).

El resto de las bandas fotométricas[4] tienen conclusiones muy similares debido a que el espectro observado no es excesivamente grande.

Es importante destacar que este análisis descriptivo es exploratorio y no controla confusores ni colinealidad. Las conclusiones se tomarán tras la realización de modelos estadísticos multivariantes evaluados con validación cruzada.

[4] Disponibles en el anexo

3.4 Tratamiento de valores atípicos

Una vez estructurados los datos se procedió al estudio de los valores atípicos (outliers). Se estudió la normalidad de las variables mediante la prueba de Shapiro-Wilk: H_0: la variable sigue una distribución normal

H_1:la variable no sigue una distribución normal.

Se obtuvieron p-valores inferiores 0.05 para todas las variables, por tanto, se concluye que las variables no están distribuidas según una distribución normal, se descartó entonces el uso de la desviación típica para el tratamiento de valores atípicos

Dado lo anterior, se optó por utilizar la desviación absoluta mediana (MAD) (Howell D. C., 2005).

MAD se presenta como un estadístico robusto ante valores atípicos al usar la mediana en vez de la media. Se define MAD como:

$$MAD_o = mediana(|X_i - mediana(X)|)$$

Ecuación 12: creación del estadístico MAD

En la literatura se usa también el MAD multiplicado por el factor 1.4826, que proviene del tercer cuartil de una distribución normal.

$$\frac{1}{\phi^{-1}(0.75)} \simeq \frac{1}{0.6745} \simeq 1.4826$$

Ecuación 13: factor de corrección del estadístico MAD

Donde $\phi^{-1}(0.75)$ es el tercer cuartil de una distribución normal estándar N(0,1).

Se declararon valores atípicos a aquellos valores que cumplieron: $|X_i - mediana(X)| \geq 6 \cdot MAD$ (operación análoga a 6σ) en cualquiera de las variables. El criterio se realizó sin conocer el estado de clasificación de la variable dependiente.

Tras la limpieza de los datos se obtuvo un data set depurado con 589 observaciones y 26 variables frente a las 827 iniciales, es decir, se eliminaron 238 observaciones.

3.5 Tratamiento de valores perdidos

3.5.1 Estudio de valores perdidos

El problema de la no respuesta es algo cotidiano en el mundo de la estadística. Este problema puede deberse a diversas causas, como a fallos en la codificación de los datos, a la pérdida de la información en el traspaso de esta, entre otras.

Históricamente se eliminaban aquellas filas que presentasen algún valor perdido, como señala Graham en *Annual Review of psychology*, esto suponía una pérdida de información ya recabada además de introducir sesgos, sobre todo si los datos se perdían de manera no aleatoria (Graham, 2009), por tanto, en los últimos años se han estudiado y presentado distintos métodos de imputación para los valores perdidos mediante modelos estadísticos, disminuyendo así el sesgo que podría generar eliminar estas observaciones y obteniendo conjuntos de datos completos (Schafer & Graham, 2002).

Para abordar correctamente los valores perdidos, es importante entender su mecanismo de ocurrencia, clasificado en tres tipos principales (Rubin D. B., 1976):

- *Missing Completely At Random* (MCAR): Se da cuando los valores perdidos, indicados por R (indicador de ausencia), no dependen ni de la parte observada de los datos (Y_{obs}) ni de los datos perdidos (Y_{miss}). Matemáticamente se puede expresar de la siguiente forma:

$$\mathrm{P(R/Y_{obs}, Y_{miss}) = P(R)}$$

Ecuación 14: Probabilidad de valor perdido bajo supuesto MCAR

- *Missing At Random* (MAR): Se da cuando los valores perdidos, R, no dependen de los propios datos perdidos (Y_{miss}) pero sí dependen de otra variable, perteneciente a la parte observada (Y_{obs}). Matemáticamente se puede expresar de la siguiente forma:

$$\mathrm{P(R/Y_{obs}, Y_{miss}) = P(R/Y_{obs})}$$

Ecuación 15: Probabilidad de valor perdido bajo supuesto MAR

- *Not Missing At Random* (NMAR): En este caso el valor perdido depende de la propia variable perdida. Matemáticamente se podría expresar de la siguiente forma:

$$\mathrm{P(R/Y_{obs}, Y_{miss})}$$

Ecuación 16: Probabilidad de valor perdido bajo supuesto NMAR

A continuación, en la ilustración 14 se muestran los patrones de los datos perdidos

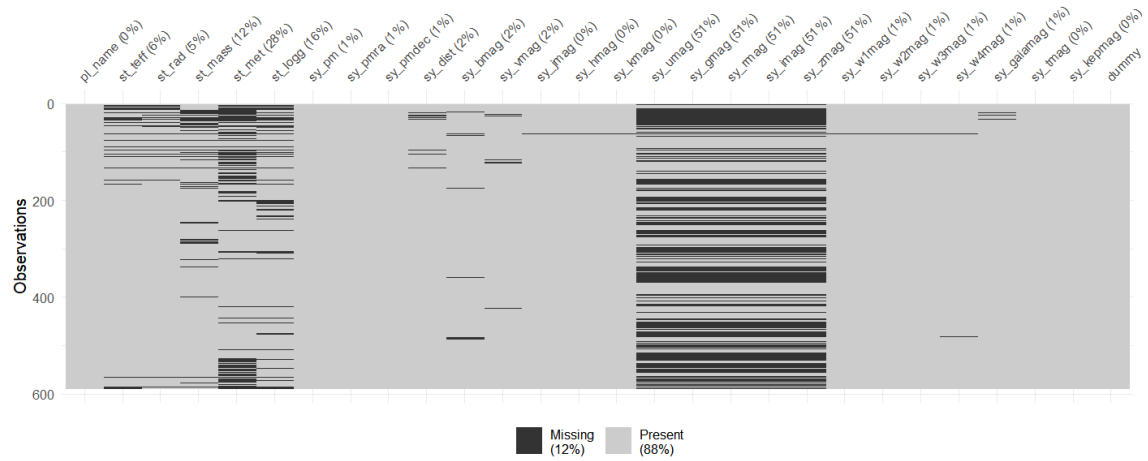


Ilustración 14: patrón de datos perdidos

Tras observar el patrón de valores perdidos, el estudio de patrones se centró en las variables problemáticas: sy_umag, sy_gmag, sy_rmag, sy_imag y sy_zmag. Parece rechazarse el supuesto MCAR, las variables sy_umag, sy_gmag, sy_rmag, sy_imag y sy_zmag presentan valores perdidos en las mismas observaciones, esto sugiere que la ausencia de estos valores está relacionada con otras variables observadas.

Para confirmar estas hipótesis se realizó la prueba de para comprobar patrón MCAR (Jamshidian, Jalal, & Jansen, 2014).

La prueba compara H_0: Los datos faltantes siguen un patrón MCAR y H_1: No hay patrón MCAR. Con un p-valor < 0.05 se rechaza la hipótesis nula sobre patrón MCAR. En consecuencia, se evalúa si la ausencia puede explicarse por las variables observadas (supuesto MAR) antes de proceder con la imputación.

Se realizó un modelo de regresión logística, con variable dependiente: los valores perdidos de la variable sy_umag. Se obtuvieron los siguientes resultados

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	1.089	2.175	0.501	0.617
sy_gaiamag	-0.099	0.107	-0.923	0.356
sy_dist	-0.002	0.001	-2.616	0.009
st_teff	0	0	0.694	0.487

Tabla 4: resultados de la logística binaria para predecir valores perdidos.

Se observó que la variable sy_dist resultó ser significativa (p-valor $= 0.009$), esto descarta patrón MCAR ya que la probabilidad de que falte un valor de sy_umag se puede modelar a partir de información observada (supuesto MAR). Teniendo en cuenta que las variables fotométricas que están siendo analizadas tienen el mismo patrón de valores perdidos, no es de extrañar, que los modelos respectivos a cada una de las demás variables reportasen el mismo resultado.

Por tanto, tras el contraste de hipótesis se rechazó el supuesto MCAR, quedando dos opciones: NMAR o MAR. Se concluyó que los valores perdidos de las variables siguen un supuesto MAR ya que pueden ser explicados por los valores observados (Y_{obs}) como se comprobó en la regresión.

3.5.2 Imputación de valores perdidos

Para la imputación de casos perdidos se ha usado el método de imputación múltiple (IM), muy recomendado en la literatura (Schafer & Graham, 2002), (Enders, Applied Missing Data Analysis, 2010). Este método de imputación es capaz de reflejar la incertidumbre derivada de los datos perdidos, gracias a hacer m distintas imputaciones. De ahora en adelante, se trabajará con m *data sets* completos, cada uno de ellos correspondiente a una imputación.

La imputación múltiple genera m > 1 data sets completos, donde cada uno presenta un posible escenario de imputación, reflejando así la incertidumbre derivada de los valores perdidos (Enders, Applied Missing Data Analysis, 2010). No existe un consenso universal sobre cuantos *data sets* son necesarios, históricamente se han recomendado 3 o 5 (Rubin D. B., 1987), aunque, como explica Enders en su libro, hay razones para usar más imputaciones ya que el error estándar disminuye según crece el número de estas (Enders, Applied Missing Data Analysis, 2010). En este trabajo se usaron m = 5 imputaciones debido a la capacidad de cómputo disponible.

Para combinar la información de los m *data sets* imputados, se aplican las reglas de Rubin que permiten obtener estimadores finales y sus varianzas ajustadas contemplando las variaciones entre imputaciones. Se define Q_{ij} como la estimación del coeficiente B_j en la imputación i = 1, …, m y U_{ij} como la varianza del coeficiente B_j en la imputación i.

$$\overline{Q_J} = \frac{1}{m}\sum_{i=1}^{m} Q_{ij}$$

Ecuación 17: Estimación combinada.

$$\overline{U_J} = \frac{1}{m}\sum_{i=1}^{m} U_{ij}$$

Ecuación 18: Varianza intra-imputación.

$$B_j = \frac{1}{m-1}\sum_{i=1}^{m}\left(Q_{ij} - \overline{Q_J}\right)^2$$

Ecuación 19: Varianza inter-imputación.

$$T_j = \overline{U_J} + \left(1 + \frac{1}{m}\right) B_j$$

Ecuación 20: Varianza total post imputaciones

$$SE_j = \sqrt{T_j}$$

Ecuación 21: Error estándar combinado

La imputación se realiza mediante un modelo de predicción, asignándole a cada valor faltante, el valor más probable según el resto de las variables. En este trabajo se han escogido modelos de *Random Forest* para la imputación (Bühlmann, 2012) debido a su capacidad de capturar relaciones complejas y su robustez frente a hipótesis de distribución. Se han entrenado *Random Forests* con 100 árboles para cada variable con valores perdidos, atendiendo a un equilibrio entre tiempo computacional y rigor. Cabe destacar que tanto la variable pl_name (identificador único para cada objeto estelar) como dummy_disposition (variable dependiente) se han excluido tanto de la lista de predictores como de la de variables a predecir. Se han obtenido finalmente 5 data sets completos con imputaciones distintas.

3.5.3 Validación de la imputación.

Es importante denotar que solo se presentaran gráficos y conclusiones para las variables con al menos un 16% de valores perdidos ya que, para el resto, al tener un porcentaje pequeño de valores perdidos, existen pocas posibilidades de ver sus distribuciones alteradas.

Durante la evaluación de las imputaciones se detectó que la variable sy_pm presentaba valores negativos en una de sus imputaciones lo que resulta imposible por su propia definición. Para corregirlo se truncaron los valores negativos a cero.

Primero se observó que las distribuciones de las variables no variasen drásticamente tras la imputación.

Como se puede observar en la Ilustración 15, se presentan las distribuciones pre-imputación (color azul) así como las distribuciones de cada imputación (color rojo) de las variables con al menos un 16% de valores perdidos. Se puede observar que las variables mantienen su distribución tras la imputación múltiple.

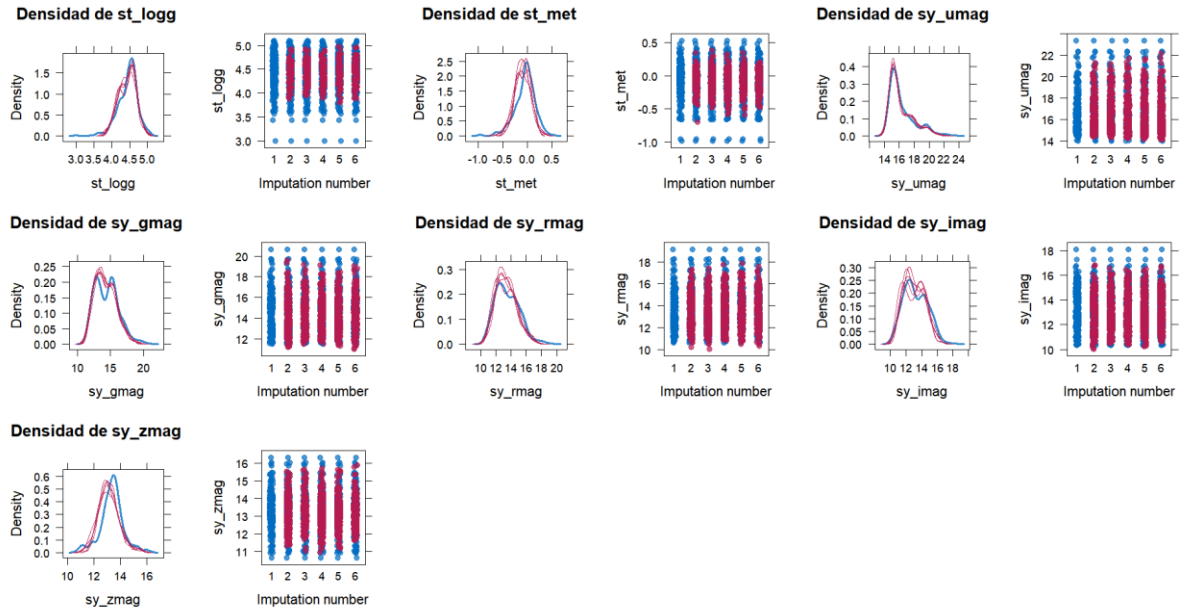


Ilustración 15: distribuciones pre y post-imputación

Para terminar de evaluar la validez del proceso de imputación se estudió la convergencia mediante trazas de medias y desviaciones típicas. En la ilustración 16 se puede observar cómo los valores de los estadísticos oscilan en torno a un nivel estable, se presentan las variables st_logg, st_met, sy_umag, sy_gmag, sy_rmag, sy_imag, sy_zmag respectivamente. En la variable sy_umag la desviación estándar muestra un incremento paulatino hasta la quinta imputación. Para futuros trabajos se recomienda aumentar el número de iteraciones con el fin de asegurar la convergencia. En este caso al ser solo una variable la que presentaba una leve incertidumbre sobre su convergencia, se prosiguió con el análisis.

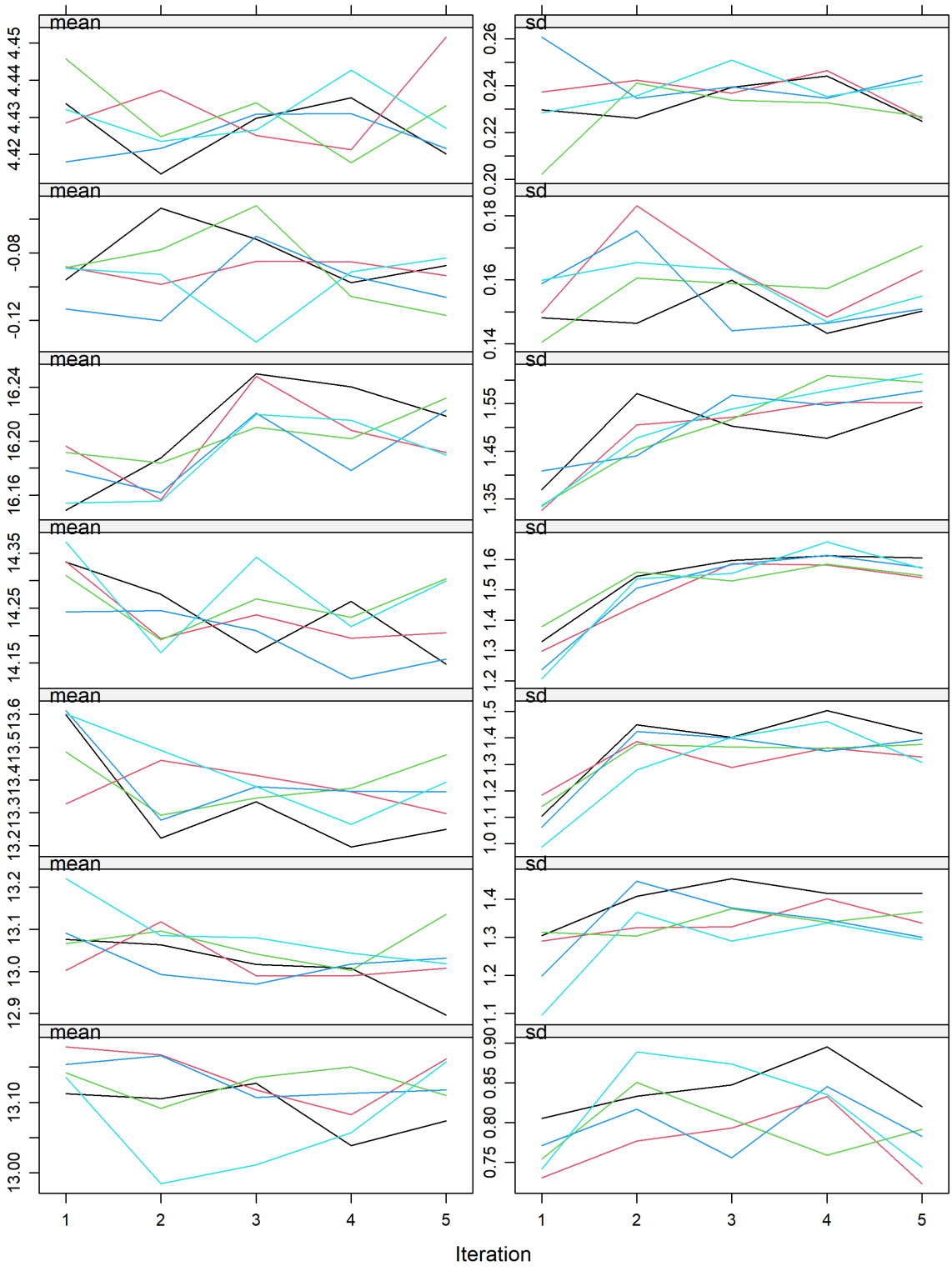


Ilustración 16: trazas de convergencia tras la imputación

4. Análisis de componentes principales

Durante el estudio se detectó que existía gran correlación entre algunas variables, en la ilustración 17 se presenta el mapa de calor:

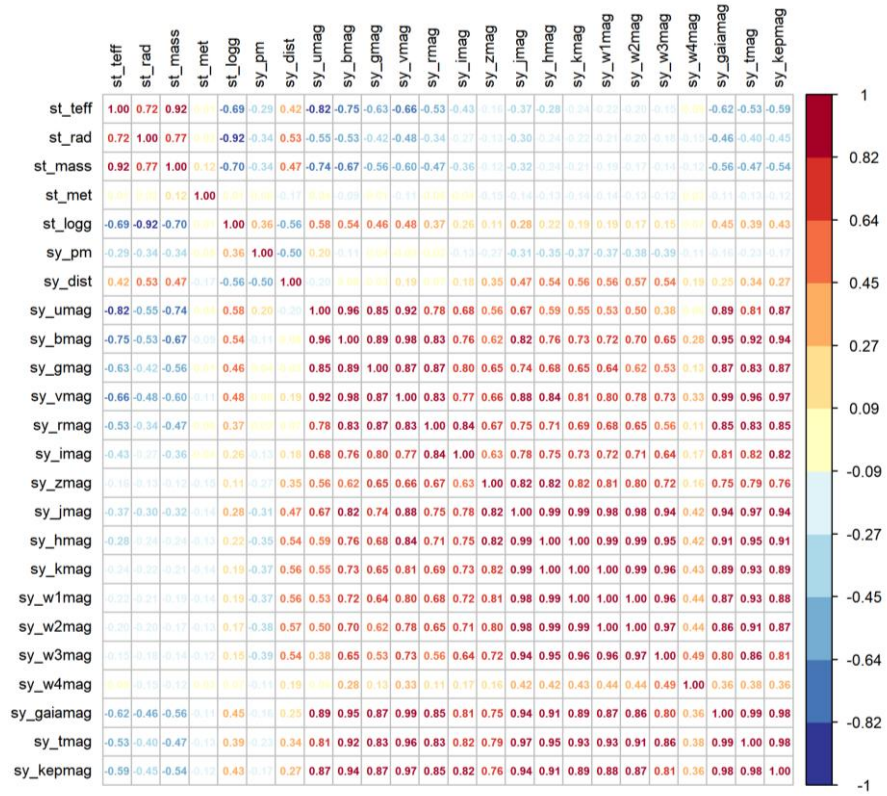


Ilustración 17: mapa de calor

Se detecta una alta correlación entre las variables fotométricas (aquellas terminadas en mag), debido a que son bandas con longitudes de onda similares, por tanto, el brillo en una de ellas se verá reflejado en las otras. También se detecta correlación entre las variables estelares (aquellas que empiezan por st).

Estas correlaciones generan un problema de multicolinealidad, esto deriva en coeficientes inestables y sesgos en las predicciones. Para solucionarlo se ha recurrido al análisis de componentes principales (ACP). De esta forma se mantiene la mayor parte de la información, además para mantener la interpretabilidad, el ACP se realizó por separado para dos grupos de variables, por un lado, las variables de carácter fotométrico y, por otro lado, las variables de parámetros estelares (Jolliffe, 2002).

Es importante destacar que se realizó el análisis en el primer *data set* imputado y se extrapolaron los resultados a los 4 restantes, de esta forma, se mantiene la misma estructura en todos los *data sets* imputados.

El número de componentes principales a retener se realizó mediante la regla de Kaiser, que consiste en seleccionar los componentes cuyos autovalores son mayores a uno (Kaiser, 1960).

4.1 ACP en variables fotométricas

Para realizar el análisis de componentes principales sobre las variables fotométricas se seleccionaron las variables sy_umag, sy_bmag, sy_gmag, sy_vmag, sy_rmag, sy_imag, sy_zmag, sy_jmag, sy_hmag, sy_kmag, sy_w1mag, sy_w2mag, sy_w3mag, sy_w4mag, sy_gaiamag, sy_tmag y sy_kepmag y se normalizaron para realizar el análisis.

En la ilustración 18, se muestra el *scree plot* con el porcentaje de varianza explicada para cada componente principal.

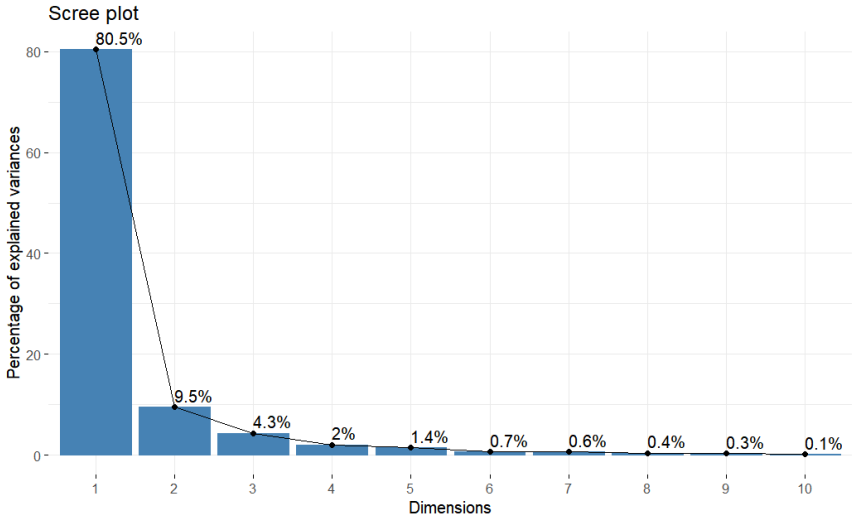


Ilustración 18: scree plot sobre los componentes de variables fotométricas

Autovalor	**Varianza Explicada**	**Varianza Acumulada**
13.67	80.46	80.46
1.61	9.51	89.97
0.73	4.34	94.31

Tabla 5: autovalores y varianza explicada de variables fotométricas

Nota. *Solo se exponen los tres primeros componentes por optimización del espacio*

De acuerdo con el criterio de Kaiser ($\lambda > 1$), se retuvieron dos componentes, explicando el 89.97% de la varianza.

A continuación, en la ilustración 19, se muestra la carga factorial de cada variable en cada componente.

En la ilustración 19 se muestran las cargas por variable en el plano (Componente1 – Componente2), se aprecia que, en la primera dimensión, todas las bandas presentan cargas de signo y magnitud similares, sugiriendo un factor común de luminosidad. Por otro lado, en la segunda dimensión, las bandas del espectro visible (U, B, V, g y r) y los filtros de banda ancha (Gaia, TESS y Kepler) muestran cargas positivas, mientras que las bandas con mayores longitudes de onda – z (borde entre espectro visible e infrarrojo cercano), J, H, K (infrarrojo cercano) y W1, W2, W3 Y W4 (borde entre infrarrojo cercano y lejano) – presentan cargas negativas, sugiriendo que el segundo factor está relacionado con el gradiente espectral o color.

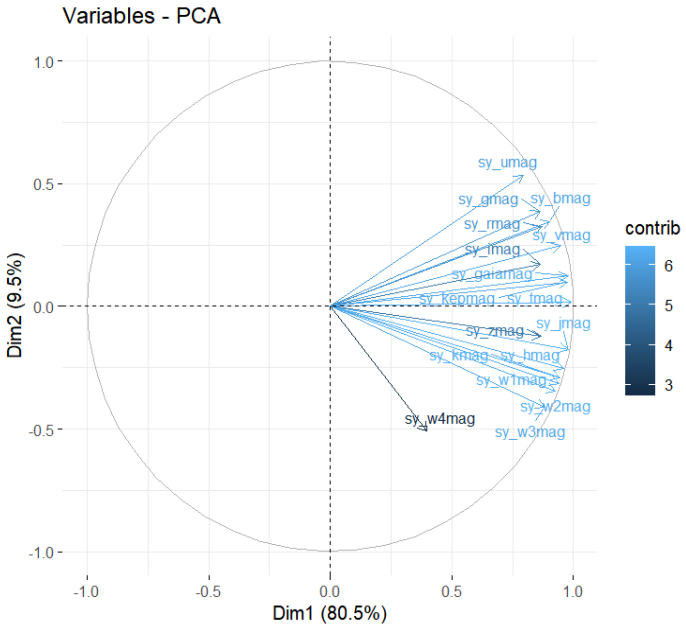


Ilustración 19: carga factorial de las variables fotométricas

4.2 ACP en variables de parámetros estelares

En este caso, se seleccionaron las variables st_teff, st_mass, st_rad y st_logg, la variable st_met se decidió dejar fuera del análisis de componentes principales por tres razones: (i) en el proceso exploratorio se obtenía una nueva componente solo para reflejar a esta variable (ii) es una variable que presenta una muy buena interpretabilidad y (iii) las variables introducidas en el análisis son de definición estructural, mientras que st_met tiene una definición más química. En la ilustración 20 se presenta el porcentaje de varianza que explica cada componente.

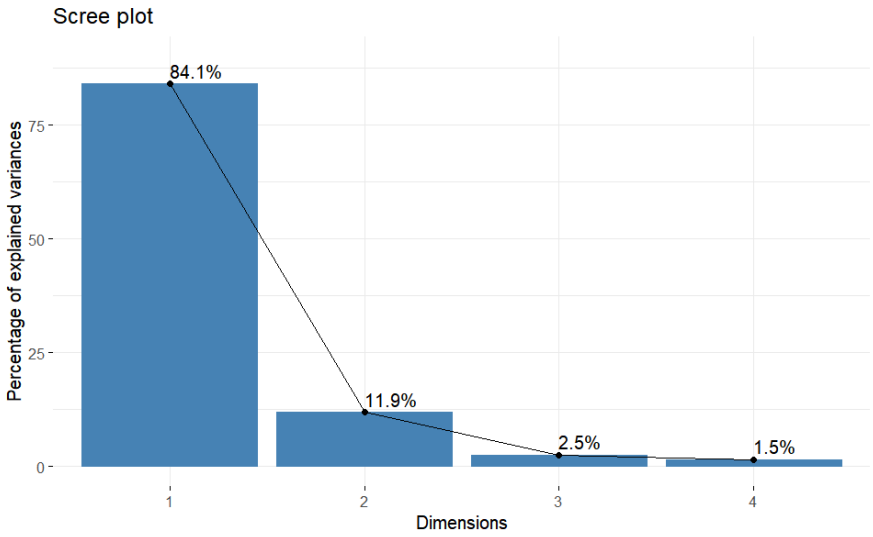


Ilustración 20: scree plot de los componentes sobre magnitudes estelares

En la tabla 6 se presenta el resumen de cada componente con los autovalores y el porcentaje de varianza de explicada. En este caso, siguiendo la regla de Kaiser ($\lambda > 1$), retendríamos un solo componente explicando así el 84.06% de la varianza explicada.

Autovalor	**Varianza Explicada**	**Varianza Acumulada**
3.36	84.06	84.06
0.48	11.91	95.97
0.1	2.52	98.48
0.06	1.52	100

Tabla 6: autovalores y varianza explicada de parámetros estelares

En la ilustración 21 se observa la contribución de cada variable a la dimensión (Componente1). Destaca que la variable st_logg es la única variable que aporta en positivo a la dimensión 1, lo que sugiere un eje de gravedad frente a tamaño, masa y temperatura. Por tanto, valores altos en este componente, indicarían que la estrella tiene alta gravedad y baja masa, temperatura y radio – estrellas compactas – mientras que valores bajos indicaran objetos con alta masa, temperatura y radio y una menor gravedad.

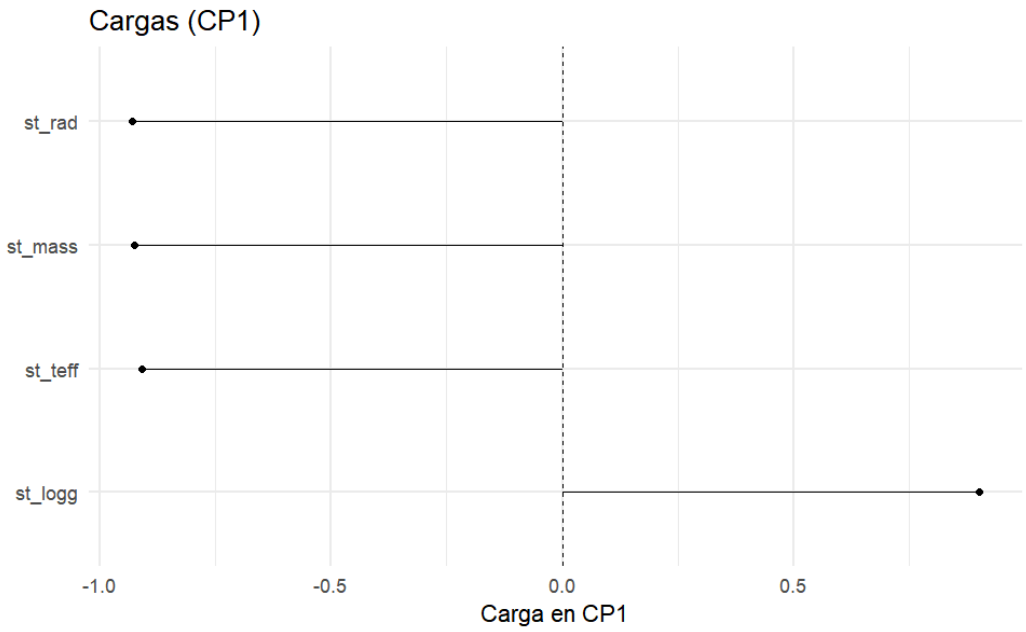


Ilustración 21: carga factorial de las variables estelares

Para evitar la multicolinealidad, se descartaron las variables originales altamente correlacionadas y en su lugar se incorporaron a la base de datos las componentes principales resultantes del ACP tanto para las variables fotométricas como para las estelares

5. Modelos básicos

5.1 Modelo de regresión logística binaria

La regresión logística es un modelo estadístico usado para predecir variables dicotómicas, es decir, variables que toman dos valores mutuamente excluyentes –en este caso, presencia o no de un exoplaneta–. Este modelo calcula la relación entre las variables independientes y la variable dependiente a partir de la función logística, que restringe la salida a 0 y 1.

Con este modelo se estudiará que parámetros estelares son más propensos a albergar la presencia de un exoplaneta y se calcularán los *odds ratio* para comprender que variables son las más influyentes.

Para mantener la fiabilidad entre imputaciones, se realizó un modelo en cada una de las 5 imputaciones y luego se combinaron los parámetros mediante el procedimiento de *pooling*, de esta forma se tiene en cuenta en los coeficientes del modelo la variabilidad entre imputaciones (Little & Rubin, 2019).

A continuación, en la tabla 7, se presentan los resultados del modelo logístico.

Estimate, es el valor medio de las estimaciones de cada imputación, se presenta también los límites inferior y superior del intervalo de confianza de cada estimador. Se observa que todas las variables son significativas al 5%.

term	***Estimate*** ($\bar{\beta}$)	***std.error***	***p-value***	$\widehat{\beta}_{2.5\%}$	$\widehat{\beta}_{97.5\%}$	***OR*** ($e^{\bar{\beta}}$)
(Intercept)	-2,4	0,501	0	-3.385	-1.415	0.09
st_met	-1.771	0,702	0,013	-3.166	-0,376	0.17
sy_pm	-0,019	0,007	0,007	-0,03	-0,005	0.98
sy_dist	0,004	0,001	0,001	0,002	0,007	1.004
PC1_mag	0,245	0,077	0,002	0,094	0,395	1.227
PC2_mag	1.094	0,136	0	0,827	1.362	2.98
PC1_st	-0,679	0,159	0	-0,99	-0,366	0.51

Tabla 7: estimadores del modelo logístico

La primera variable, st_met, tiene un coeficiente medio estimado negativo, lo que implica que, controlando el resto de predictores, planetas que orbiten estrellas más metálicas tienen una menor probabilidad de ser clasificados como exoplanetas. En términos de *odds ratio* (OR) se obtiene un OR medio de 0.17 (IC95% (0.04, 0.69)), es decir, por cada aumento de dex, disminuyen en 83% las razones de probabilidad (*odds*) de catalogar al objeto como exoplaneta. Esta conclusión es sorprendente, ya que toma la dirección contraria al análisis bivariado anterior, en el que se concluyó que, a mayor metalicidad, mayor proporción de exoplanetas, como se nombró anteriormente, esto se debe a que el análisis bivariado tiene una función exploratoria y no tiene en cuenta confusores. Por otro lado, en la literatura consultada, se encuentra una relación positiva entre metalicidad y la confirmación de gigantes gaseosos (Fischer, 2008). Esta conclusión invertida se obtiene debido a que solo el 3% de la muestra de exoplanetas usada para la creación del modelo está catalogada como gigante gaseoso.

Continuamos con la variable sy_pm, con un coeficiente medio estimado negativo y cercano a cero, aunque significativo. Se obtiene un OR medio de 0.98 (IC95% (0.96, 0.99), por tanto, un aumento de la velocidad aparente de un milisegundo de arco por año en la estrella anfitriona reduce en 0.98 las *odds* de ser catalogado como exoplaneta al objeto que la orbite.

La variable sy_dist presenta también un coeficiente positivo y prácticamente nulo, aunque significativo también, por tanto, al aumentar la distancia, aumentan las *odds* de catalogar al

objeto como exoplaneta. Con un OR medio de 1.004 (IC95% (1.001, 1.006)), un aumento de un parsec al sistema planetario aumenta en 1.004 veces las razones de probabilidad de que el objeto estelar sea catalogado como exoplaneta.

A continuación, se presenta la variable PC1_mag, primer componente principal de las variables fotométricas, definido como brillo global. Con un coeficiente medio estimado de 0.245 se concluye que, mayores magnitudes (brillo más tenue) en las bandas fotométricas, existe una mayor posibilidad de catalogar el objeto como exoplaneta. En termino de *odds*, se ha obtenido un OR medio de 1.277 (IC95% (1.1, 1.48)), por tanto, por cada aumento en la magnitud (*score*) de esta variable, se tienen 1.277 veces más *odds* de catalogar al objeto estelar como exoplaneta.

La variable PC2_mag, segundo componente principal de las variables fotométricas, definido como gradiente espectral. Presenta un coeficiente medio estimado de 1.094 (IC95% (0.827, 1.362)), por tanto, un crecimiento en el *score* de esta variable implica un aumento en la probabilidad de catalogar al objeto como exoplaneta. Se obtuvo un OR de 2.98, por cada unidad que aumente PC2_mag, se tienen prácticamente 3 veces más *odds* de catalogar al objeto como exoplaneta, es decir, un espectro relativamente más dominado por el óptico visible (véase la ilustración 19 de cargas factoriales), prácticamente triplica las *odds* de catalogar el objeto como exoplaneta.

Por último, la variable PC1_st, primer y único componente del grupo de parámetros estelares, definido como un diferenciador de estrellas compactas o no. Con un coeficiente medio estimado de -0.679 tiene una relación inversa con la variable dependiente, por tanto, el aumento de esta variable – estrellas compactas – indican una menor probabilidad de catalogar al objeto como exoplaneta. En términos de *odds ratio* se obtuvo un valor de 0.51 (IC95% (0.37, 0.69)), es decir, por cada *score* que aumente la variable, disminuye en 0.51 las posibilidades de catalogar al objeto como exoplaneta.

5.1.1 Evaluación del modelo

El rendimiento del modelo se evaluó mediante validación cruzada (CV) con k = 5 particiones y agrupado por sistema (host_id[5]). La agrupación se realizó para evitar fuga de información entre objetos del mismo sistema que pudiesen caer en diferentes pliegues (*folds)*. Se definió host_id como identificador del sistema. Se usaron cinco *folds* y se usó la misma partición en cada una de las cinco imputaciones.

[5] host_id, proviene de quitar el identificador único a cada uno de los nombres de pl_name

Para cada observación se almacenaron las probabilidades *out of fold* (OOF) de cada imputación (p_{ik}), se promediaron estas probabilidades entre las m = 5 imputaciones, como se muestra en la ecuación 22.

$$\bar{p}_i = \frac{1}{m}\sum_{k=1}^{m} p_{ik}$$

Ecuación 22: promedio de probabilidades OOF

Finalmente se calculó el área bajo la curva ROC (AUC) con las probabilidades promediadas entre imputaciones, obteniendo un AUC = 0.875, valor que indica una buena capacidad de discriminación.

En la ilustración 22 se presenta cada curva ROC por imputación, así como la curva ROC final.

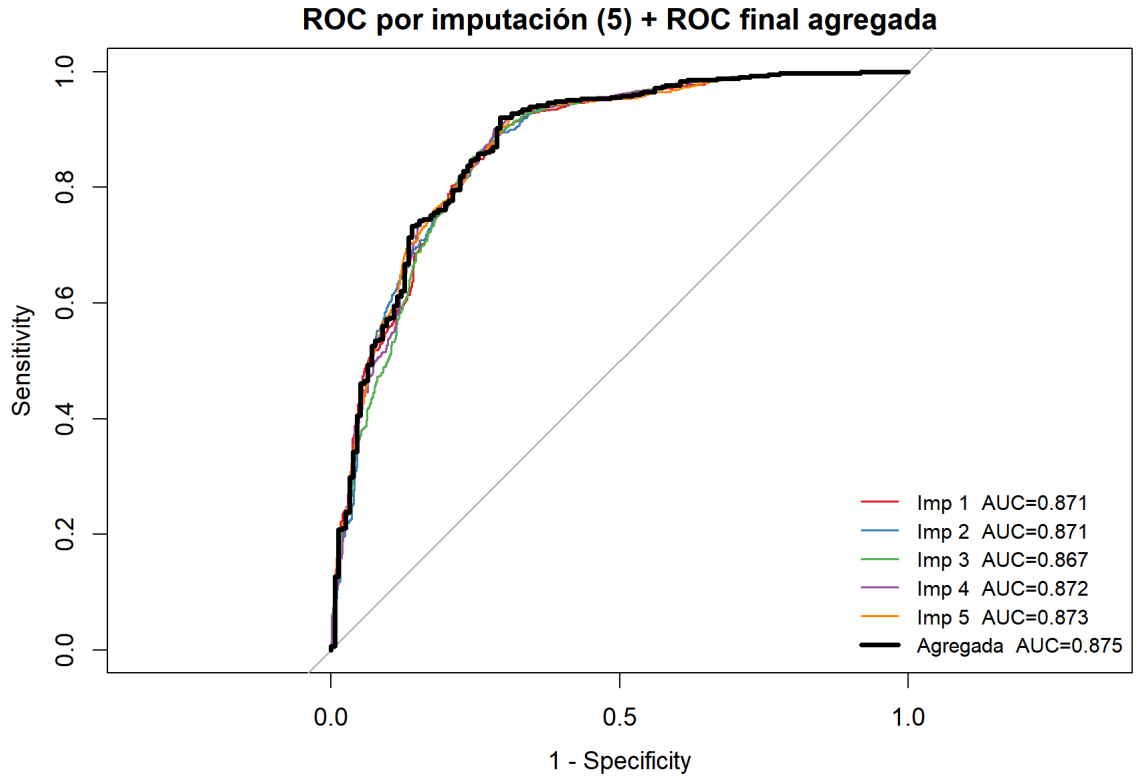


Ilustración 22: curva ROC logística binaria

Es importante destacar que el AUC no tiene en cuenta la prevalencia de las clases, para tener información sobre la calidad de discriminación de éxitos (exoplaneta) y fracasos (no exoplanetas) se calculó la sensibilidad, especificidad y tasa de acierto para el punto de corte 0.5

Umbral	Sensibilidad	Especificidad	*Accuracy*	*Balanced Accuracy*	PPV	NPV
0.5	0.942	0.643	0.863	0.7925	0.88	0.8

Tabla 8: evaluación del modelo logístico punto de corte 0.5

A continuación, en la tabla 9 se presenta la matriz de confusión.

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	101	56
Exoplaneta (observado)	25	407

Tabla 9: matriz de confusión para el punto de corte 0.5

No es de extrañar la gran diferencia que se encuentra entre sensibilidad y especificidad, ya que la variable dependiente esta desbalanceada. El modelo fue capaz de catalogar correctamente al 94.21% de los objetos clasificados como exoplanetas, por otro lado, el desempeño clasificando a los no exoplanetas, no fue tan positivo, tan solo el 64.33% de los objetos fueron clasificados como no exoplanetas cuando realmente no lo eran. En general la tasa de acierto fue de 86.3 % de los objetos estelares, pero la tasa de acierto balanceada (*balanced accuracy*) fue del 79.3%, esta diferencia se debe al desbalanceo de clase, con la tasa de acierto balanceada se le da el mismo peso a la sensibilidad y especificidad.

En términos de valor predictivo, cuando el modelo predice “exoplaneta” acierta en el 88% de las veces, cuando el modelo predice no exoplaneta acierta en el 80%.

A continuación, en la tabla 10 y 11, se presentan las métricas de validación y la matriz de confusión respectivamente, para el punto de corte óptimo de Youden:

Umbral	Sensibilidad	Especificidad	*Accuracy*	*Balanced Accuracy*	PPV	NPV
0.587	0.921	0.7	0.864	0.81	0.896	0.765

Tabla 10: evaluación del modelo logístico punto de corte óptimo

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	111	46
Exoplaneta (observado)	34	398

Tabla 11: matriz de confusión para el punto de corte óptimo

No se encuentran grandes mejoras entre los dos puntos de corte, debido a que el punto óptimo de Youden es muy cercano al punto por defecto. De todas formas, se observa una mejora de prácticamente 6 puntos porcentuales en la especificidad, a costa de un dos por ciento de sensibilidad. Ambas tasas de acierto (tanto la habitual como la balanceada) presentan una leve mejora.

5.2 Árbol de clasificación

Para construir un árbol de clasificación, existen diferentes criterios de división que pueden dar lugar a distintos resultados. En este caso se probará con el índice de Gini y la entropía.

Se entrenaron dos árboles, cada uno de ellos con un criterio de división, pero bajo la misma validación cruzada e idénticos *folds* en las cinco imputaciones. Se obtuvo un AUC_{gini} = 0.796 y un $AUC_{entropía} = 0.81$; se realizó la prueba de DeLong pareada para contrastar si existen diferencias entre las áreas bajo las curvas ROCs (Clarke-Pearson, DeLong, & DeLong, 1988). Consiste en un test no paramétrico que compara AUCs cuando estas están correlacionadas. Se obtuvo un p-valor = 0.1729, por tanto, no se han encontrado evidencias suficientes en contra de la hipótesis nula de igualdad de AUCs.

En la construcción del árbol de decisión se decidió ajustar el mínimo tamaño de hoja a 20, de esta forma, se evitó crear un árbol muy complejo y garantizar así la interpretabilidad. Además, se tuneó el parámetro de complejidad, que define el mínimo beneficio que debe aportar una división para realizarse con valores desde 0.001 hasta 0.02 y se seleccionó el criterio de entropía para la división de nodos.

En la ilustaración 23, donde se presenta la evolución del AUC según el *complexity parameter* en cada imputación, se puede observar cómo según aumenta el cp (menor complejidad) disminuye el AUC. Para todas las imputaciones se escogió cp = 0.007, excepto para la quinta imputación, en la cual se seleccionó un cp = 0.015.

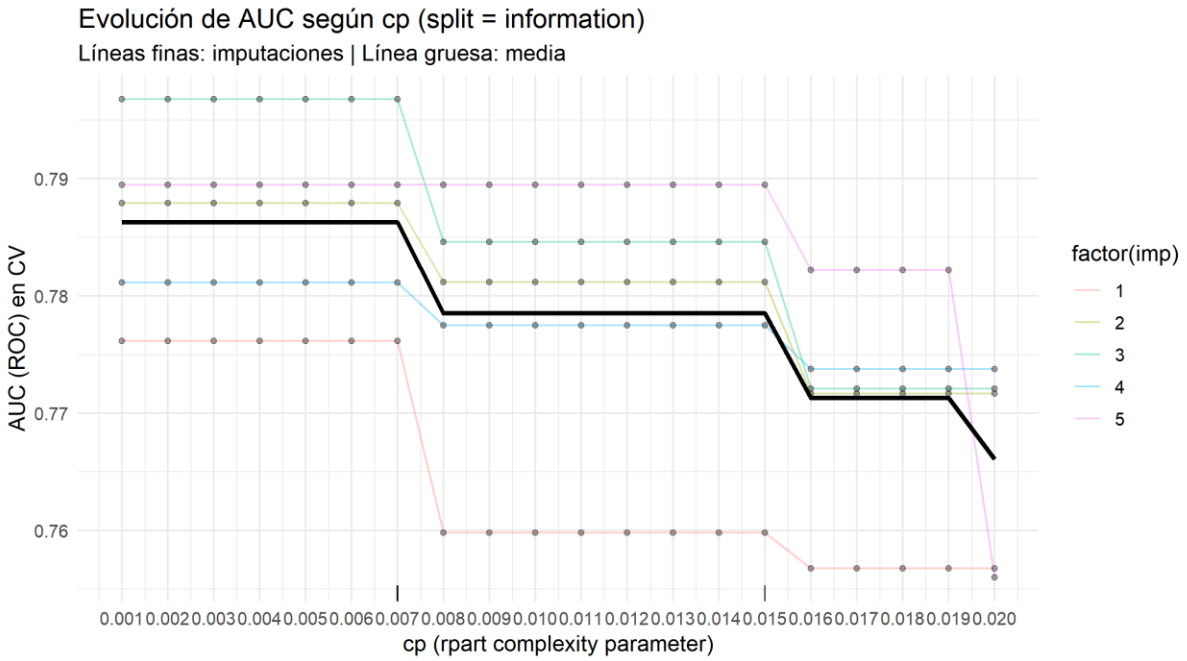


Ilustración 23: evolución de AUC, según cp

Para la creación de los árboles no se usaron los componentes creados anteriormente, por dos motivos, 1) los árboles son más robustos frente a multicolinealidad y 2) se buscan perfiles interpretables y los componentes principales restan interpretabilidad al modelo.

En la ilustración 24 se presenta el diagrama de árbol generado.

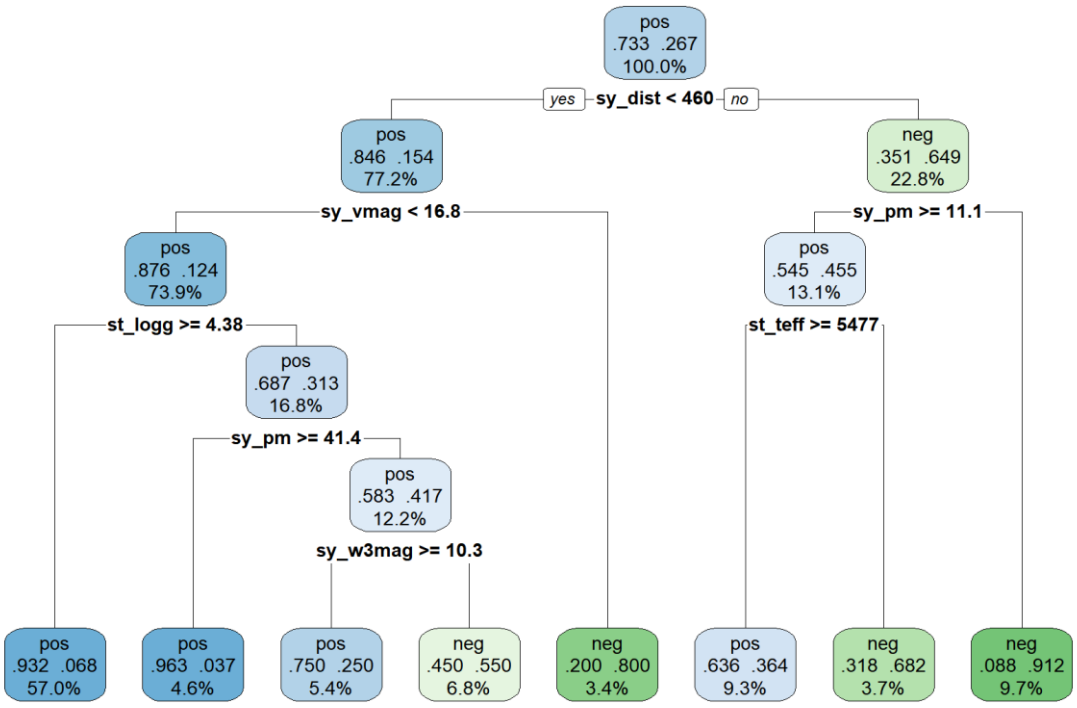


Ilustración 24: árbol CART, regls y clase predicha por nodo

Con el diagrama de árbol es fácil generar un perfil que resuma que sistemas son más propensos a albergar objetos con próxima clasificación de exoplanetas. Se observa, que aquellos objetos que se alejen menos de 460 parsecs de la estrella anfitriona, que esta tenga una magnitud en la banda V menor a 16.8 y la aceleración de la estrella se mayor a 4.38 cm/s^2, tienen una alta probabilidad de ser clasificados como exoplanetas (P(Exoplaneta) = 0.932). Así mismo, en el subgrupo generado por sy_dist < 460 y sy_vmag < 16.8, si la aceleración de la estrella es mayor a 4.38 cm/s^2 y la velocidad angular mayor a 41.4 milisegundos de arco por año, el objeto tiene una probabilidad de 0.963 de ser clasificado como exoplaneta. Por otro lado, los objetos que se alejen más de 460 parsecs de la estrella anfitriona y la velocidad angular de esta sea inferior a 11.1 milisegundos de arco por año tienen una alta probabilidad de ser clasificados como no exoplanetas, P(No Exoplaneta) = 0.912.

En la ilustración 25 se presenta el árbol junto a un gráfico de barras con la proporción de la variable dependiente en cada hoja.

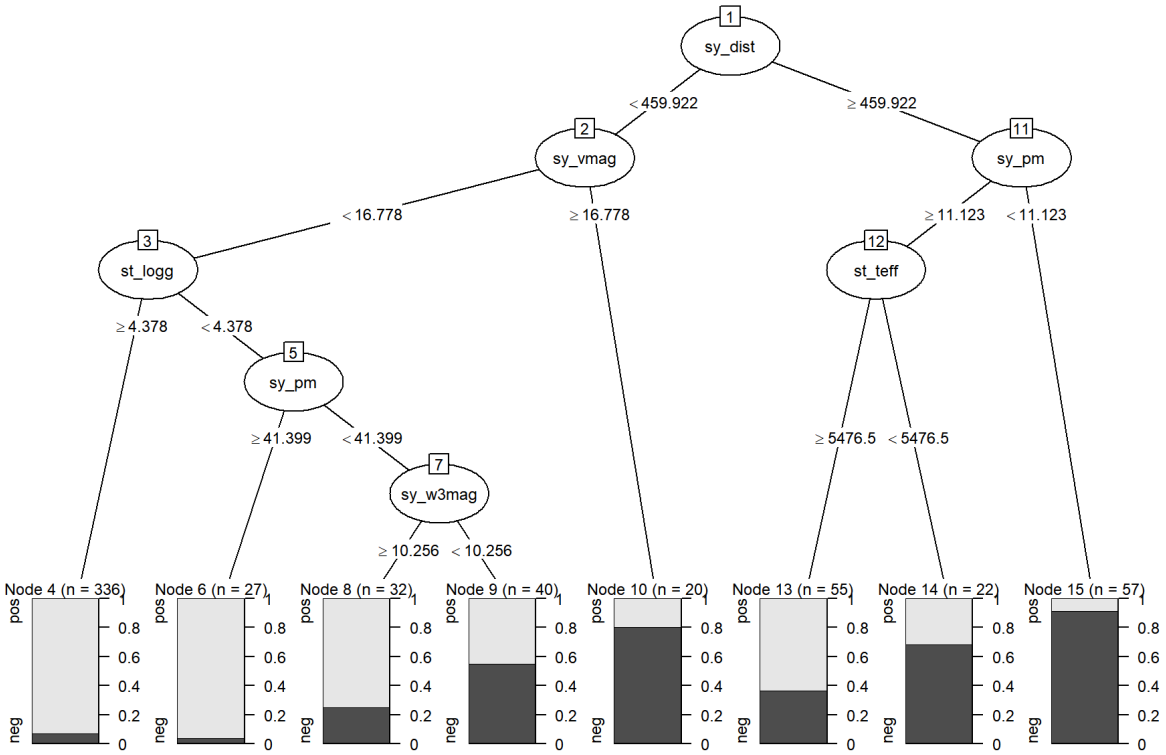


Ilustración 25: pureza y tamaño muestral por hoja

Este diagrama de árbol es realmente útil para entender la puridad de las hojas. Se nombran las hojas de izquierda a derecha empezando por el uno y terminado por el ocho.

La hoja más a la izquierda (hoja número 1) es la de mayor tamaño muestral y está formada por prácticamente todo exoplanetas, por tanto, es una hoja fiable. En la misma línea, la hoja número 2, aunque tiene una menor proporción, prácticamente es una hoja totalmente pura, lo que la hace fiable. La hoja más a la derecha (hoja número 8) presenta una proporción de casos similar, pero a la inversa, y aunque tenga un tamaño de hoja menor, se considera una hoja fiable.

Las hojas 4, 6 y 7 son hojas de alta incertidumbre, la proporción de casos está prácticamente igualada y tienen pocas observaciones. Las conclusiones que se puedan sacar sobre estos nodos no deberían de ser muy fiables.

5.2.1 Evaluación del modelo

Se entrenó un árbol mediante validación cruzada 5-fold y estratificada por host_id en cada una de las imputaciones, se usó el área bajo la curva ROC como medida de selección del mejor modelo, además se almacenaron las predicciones OOF para el futuro cálculo de la sensibilidad, PPV, especificidad, NPV, tasa de acierto e índice Kappa, el cual toma valores entre 0 equivalente a que el modelo clasifica al azar y 1 equivalente a que el modelo clasifica perfectamente. El cálculo de Kappa se presenta a continuación:

$$k = \frac{P_o - P_e}{1 - P_e}$$

Ecuación 23: cálculo de índice Kappa

Siendo $P_o = \frac{VP+VN}{N}$ y $P_e = \frac{(VP+FP)(VP+FN)+(FN+VN)(FP+VN)}{N^2}$, N es el número total de observaciones. Para valorar el índice se usará la escala de Landis & Koch, en la que a partir de un índice de 0.41 se considera un desempeño moderado, y a partir de 0.61 un buen desempeño (Landis & Koch, 1977)

El mejor árbol reportó un AUC de 0.825, a continuación, en la tabla 12 se presenta el cálculo de las métricas de evaluación para el punto de corte por defecto (0.5) y en la tabla 12 se presenta la matriz de confusión.

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.5	0.946	0.433	0.82	0.74	0.81	0.69	0.438

Tabla 12: evaluación del árbol para el punto de corte 0.5

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	68	89
Exoplaneta (observado)	23	409

Tabla 13: matriz de confusión umbral 0.5

Es notable la gran disparidad entre sensibilidad (0.946) y especificidad (0.433), como se explica en el modelo anterior, esto se debe a la prevalencia elevada de la clase positiva (Exoplaneta) además el punto de corte 0.5 favorece esta situación, esto explica también porque existe una diferencia de 12 puntos porcentuales entre la tasa de acierto y la tasa de acierto balanceada.

El modelo tiene un desempeño moderado según el índice Kappa y en términos de valores predictivos, de los objetos catalogados como exoplanetas el 82% lo fueron realmente y de los objetos catalogados como no exoplanetas el 74% no lo fueron realmente.

A continuación, en la tabla 14 se presenta las métricas para el umbral óptimo de Youden y en la tabla 15 la matriz de confusión.

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.84	0.768	0.783	0.9	0.55	0.77	0.77	0.48

Tabla 14: evaluación del árbol para el punto de corte de Youden

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	123	34
Exoplaneta (observado)	100	332

Tabla 15: matriz de confusión para el umbral de Youden

Con el umbral óptimo de Youden se obtiene una mejora de 35 puntos porcentuales en especificidad, consiguiendo detectar correctamente al 78.3% de los no exoplanetas presentes en la base de datos, a cambio de restarle 17.8 puntos a la sensibilidad, respecto al punto de corte 0.5. En términos de equilibrio se nota una clara mejora del árbol de clasificación. Ambas

tasas de acierto se han mantenido relativamente constantes, en términos predictivos, PPV ha incrementado en ocho puntos porcentuales (menos falsos positivos), mientas que NPV ha decrecido en 19 puntos porcentuales (más falsos negativos).

Por tanto, si el objetivo a conseguir es maximizar aciertos, entonces se ha de optar por el modelo con el umbral óptimo de Youden, en caso de que no se quiera perder ningún exoplaneta, el modelo con el punto de corte por defecto tendría un mejor desempeño. Nótese que PPV/NPV dependen de la prevalencia del conjunto evaluado; si la prevalencia real difiere de la del *data set*, estos valores cambiarán.

6. Modelos avanzados

6.1 *Random Forest*

Tras la regresión logística y el árbol de decisión, se ha realizado un modelo de *Random Forest* (RF) con la intención de mejorar la capacidad predictiva del análisis.

El algoritmo de *Random Forest* (Breiman, 2001), mejora la capacidad predictiva de los árboles añadiendo dos fuentes claves de variabilidad:

1. Genera muestras aleatorias con reemplazamiento (*bootstrap*) y para cada muestra construye un árbol independiente.
2. En lugar de evaluar todas las variables del conjunto de datos en cada árbol, lo que daría lugar a un modelo *Bagging*, se escoge un número aleatorio de variables para la creación de cada árbol, esto permite que variables con una calidad predictiva media no queden opacadas por aquellas con una calidad alta.

La desventaja de los modelos de RF es que no se pueden visualizar debido al gran número de árboles que generan, para añadir calidad interpretativa, se calculó la importancia de cada variable mediante la permutación de variables, en la que se mide el incremento del error al eliminar una variable del modelo.

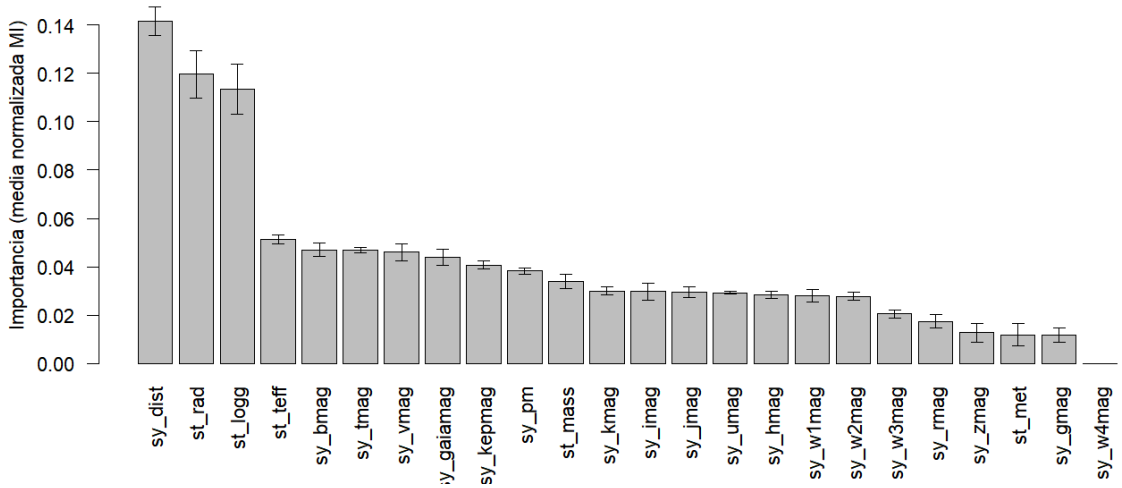


Ilustración 26: importancia de variables de RF

Como se observa en la ilustración 26, la variable más importante es sy_dist, seguida por st_rad y st_logg, que destacan considerablemente comparadas con el resto de las variables.

El modelo RF se creó con 1000 árboles en cada imputación, el desbalanceo se trató con pesos de clase, a la clase minoritaria se le asignó un peso contrario a su prevalencia. El rendimiento del modelo se evaluó mediante las observaciones OOF. El modelo se entrenó mediante validación cruzada 10-fold y agrupada, se seleccionaron los hiperparámetros (desarrollados en el marco teórico) según maximizasen el AUC, en la ilustración 27 se presenta la evolución del AUC según los hiperparámetros. La combinación óptima, que coincidió en cada imputación, fue mtry = 10, splitrule = extratrees y min.node.size = 1

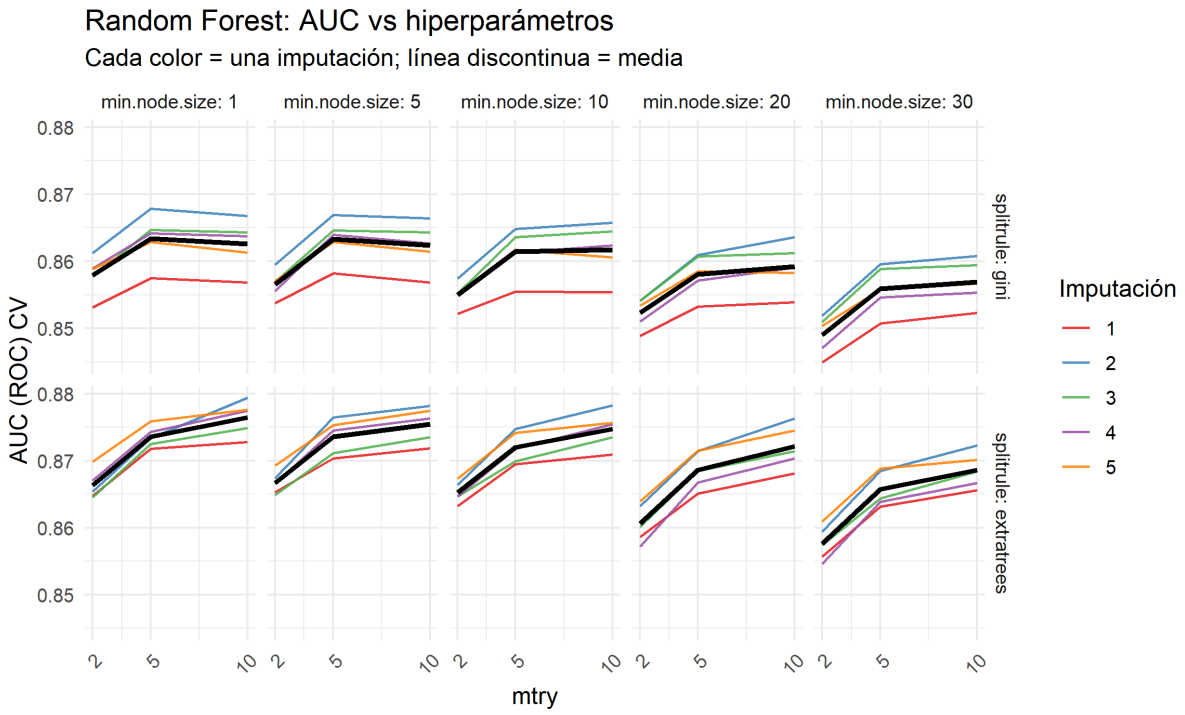


Ilustración 27: evolución de hiperparámetros de RF

Se puede observar como la fila de *extratrees* presenta AUCs levemente superiores en todas las imputaciones, así como el aumento de mtry parece aumentar el AUC.

6.1.1 Evaluación del modelo

Se obtuvo un área bajo la curva ROC de 0.88, en la tabla 16 se presentan las métricas de evaluación para el punto de corte 0.5

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.5	0.935	0.586	0.86	0.76	0.84	0.76	0.56

Tabla 16: evaluación de RF, punto de corte 0.5

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	92	65
Exoplaneta (observado)	28	404

Tabla 17: matriz de confusión RF, punto de corte 0.5

Se obtiene una sensibilidad de 0.935, es decir, se detectaron el 93.5% de los exoplanetas presentes en la base datos. Por otro lado, se obtiene una especificidad de 0.586, por lo que el modelo solo detectó el 58.6% de los no exoplanetas presentes en la base de datos. En términos predictivos se obtiene un PPV de 0.86, es decir, de los objetos predichos como exoplaneta el 86% lo fueron realmente. Para NPV se obtiene un valor de 0.76, por lo que de los objetos clasificados como no exoplanetas, realmente el 76% no lo fueron.

El modelo ha obtenido una tasa de acierto del 84%, aunque disminuye en un 76% para la tasa de abierto balanceada (que no tiene en cuenta la clase minoritaria) notamos un descenso de ocho puntos porcentuales.

En cuando al índice Kappa, con un valor 0.56 se concluye que el modelo tiene un desempeño moderado.

A continuación, se presenta la evaluación para el punto de corte óptimo de Youden, en el que se busca maximizar sensibilidad y especificidad.

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.64	0.87	0.75	0.9	0.68	0.84	0.81	0.6

Tabla 18: evaluación de RF, umbral de Youden

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	118	55
Exoplaneta (observado)	39	377

Tabla 19: matriz de confusión RF, umbral de Youden

Usando el umbral de Youden, se obtiene una mejora significativa en la especificidad del modelo, pasando del 58.6% al 75% (incremento de 16.4 puntos porcentuales), este incremento en la especificidad se consigue a costa de 6.5 puntos porcentuales en la sensibilidad. En cuanto términos predictivos, se observa una ligera mejora en el valor predictivo positivo, en cuanto al valor predictivo negativo se obtiene un decremento de 8 puntos porcentuales.

En términos generales de clasificación, destaca el aumento de la tasa de acierto balanceada llegando al 81% y con un índice Kappa de 0.6 se puede clasificar al modelo como bueno.

6.2 *Extreme Gradient Boosting*

Con intención de comparar distintos modelos con ciertas similitudes, se entrenó un modelo de boosting, se espera que XGBoost mejore el AUC frente a la regresión logística y el árbol de clasificación, así como que compita con el *Random Forest*.

El tuneo de hiperparámetros (desarrollado en el marco teórico) se hizo en dos fases para disminuir el tiempo de cómputo. En una primera fase, se usó una rejilla reducida para localizar la zona óptima donde probar, la combinación que maximizó el AUC en primera instancia fue la siguiente: *nrounds* 400, *maxdepth* 4, η 0,1, *gamma* 0, *colsample_bytree* 0.8, *min_child_weight* 3, *subsample* 0.8

Para la segunda fase se volvieron a tunear los parámetros:

- nround, multiplicándolo por 0.75, 1 y 1.25
- max_depth, restándole 1, sumándole 0 y 1
- gamma se probó con 0 y 1
- min_child_weight se probó con 1, 2 y 3

finalmente, la combinación que maximizó el AUC en cada una de las imputaciones fue:

Imputación	***nrouds***	***Max_depth***	**η**	**Gamma**	***Colsample by tree***	***Min child weight***	***subsample***
1	400	5	0.1	0	0.8	3	0.8
2	500	3	0.1	1	0.8	2	0.8
3	500	5	0.1	0	0.8	1	0.8
4	300	5	0.1	0	0.8	2	0.8
5	400	4	0.1	0	0.8	2	0.8

Tabla 20: selección de hiperparámetros en XGboost

En la ilustración 28 se presenta la importancia de las variables, podemos ver que la variable más importante es sy_dist, seguida por st_logg, sy_pm, st_teff y st_rad.

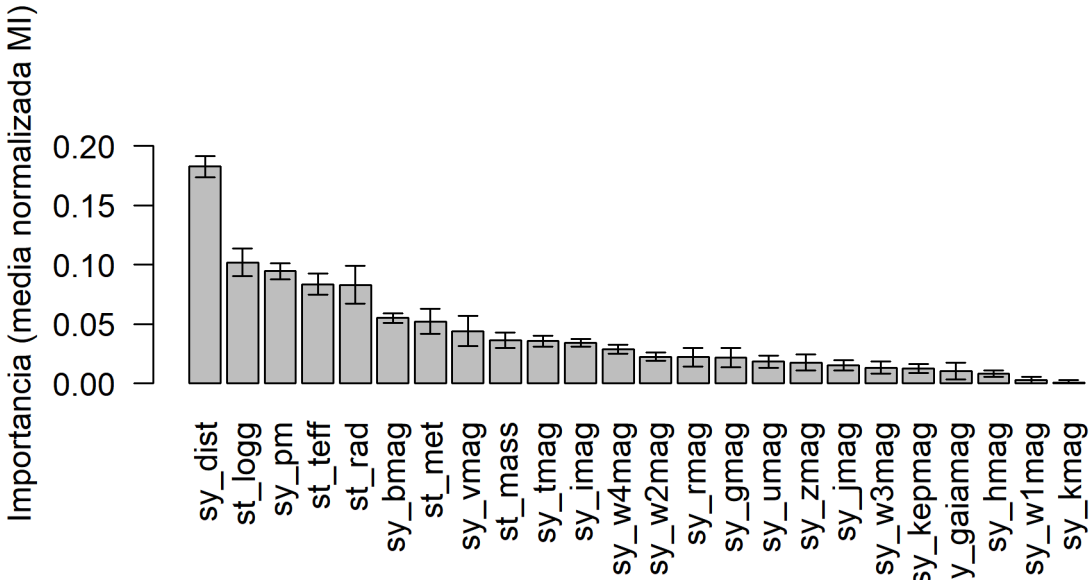


Ilustración 28: importancia de variable en XGBoost

6.2.1 Evaluación del modelo

Se obtuvo un área bajo la curva ROC de 0.88, en la tabla 21 se presentan las métricas de evaluación para el punto de corte 0.5

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.5	0.94	0.63	0.875	0.792	0.857	0.785	0.61

Tabla 21: evaluación de XGBoost, umbral 0.5

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	99	58
Exoplaneta (observado)	26	406

Tabla 22: matriz de confusión XGBoost, umbral 0.5

La evaluación del modelo es bastante positiva, con una sensibilidad de 0.94, del total de exoplanetas en la base de datos, se detectaron correctamente al 94%, en cambio se obtuvo una especificidad de 0.63; es decir, solo se detectaron el 63% de los no exoplanetas presentes en la base de datos. Los valores predictivos alcanzan niveles óptimos, con un valor predictivo positivo que indica que de los objetos clasificados como exoplanetas el 87.5% realmente lo

fueron. El valor predictivo negativo alcanza un valor de 0.792, es decir, de los objetos clasificados como no exoplanetas el 79.2% realmente no lo fueron.

En términos de predicción global se obtiene una tasa de acierto de 0.857; sin embargo, la tasa de acierto balanceada baja hasta 0.785. Además, el índice Kappa de 0.61 indica una buena calidad del modelo.

Se presenta a continuación las métricas para el punto de corte óptimo de Youden

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	***Accuracy***	**Balanced Accuracy**	**Kappa**
0.82	0.87	0.82	0.93	0.70	0.859	0.8472	0.65

Tabla 23: evaluación de XGBoost, umbral óptimo de Youden

	No exoplaneta (pronosticado)	**Exoplaneta (pronosticado)**
No exoplaneta (observado)	129	28
Exoplaneta (observado)	55	377

Tabla 24: matriz de confusión XGBoost, umbral óptimo de Youden

Los resultados con el punto de corte óptimo de Youden muestran una mejora sustancial. La especificidad experimenta un aumento de diecinueve puntos porcentuales mientras que la sensibilidad solo decrece en siete puntos porcentuales. En cuanto al valor predictivo positivo se experimenta una leve mejora y por otro lado el valor predictivo negativo presenta un decrecimiento de nueve puntos porcentuales.

En términos de calidad global, la tasa de fallo no presenta grandes cambios, en cambio, sí se observa una mejora de seis puntos porcentuales para la tasa de fallo balanceada. Lo que indica una gran capacidad predictiva.

6.3 Comparación de modelos

Con intención de resumir la comparación de modelos, en la ilustración 29, se presentan unos diagramas de cajas y bigotes para las tres métricas más importantes, AUC, sensibilidad y especificidad.

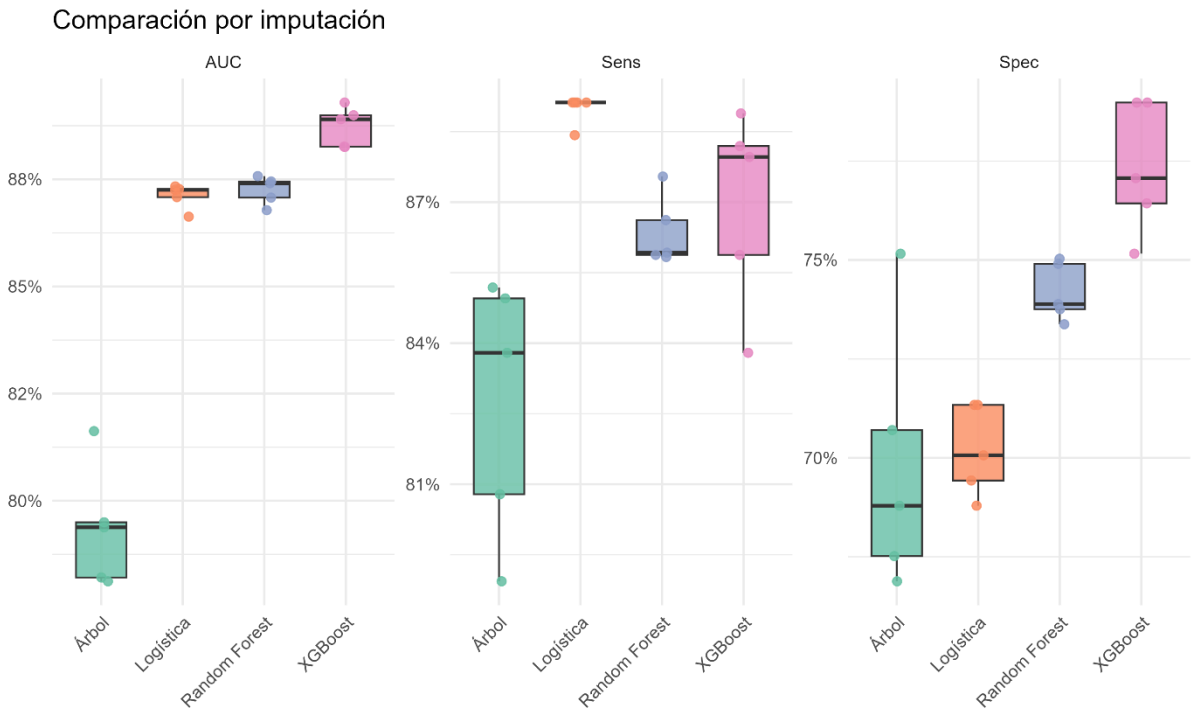


Ilustración 29: Comparación de modelos

Observando el gráfico se concluye que el árbol de clasificación es el modelo con peores métricas e inclusive mayor variación entre imputaciones (coherente con la alta varianza típica de un solo árbol), no se puede clasificar como mal modelo, pero sí que es peor que el resto. En cuanto a la regresión logística, destaca que es el modelo con menor variabilidad, comparable a la del modelo *Random Forest*. Por último, el modelo con mejor desemepeño es XGBoost, aunque tiene una variabilidad superior que la regresión logística y RF, toma valores visiblemente superiores sobre todo en cuanto a especificidad.

En conclusión, los modelos clásicos son útiles gracias a su posible interpretación de parámetros, pero en búsqueda de una buena clasificación destacan los modelos de caja negra.

7. Limitaciones y líneas futuras

Limitaciones. Este estudio utiliza los datos de la misión K2, esto podría introducir sesgos de selección y dominio no generalizables a otros catálogos. La variable dependiente presenta un desbalance de clases que dificulta que el AUC refleje de manera adecuada la calidad de los modelos. En el conjunto K2 existen grupos naturales de exoplanetas que orbitan a la misma estrella, aunque esto se tuvo en cuenta en la validación cruzada, en los modelos se usó cada planeta como independiente. El *pooling* ente imputaciones realizado para estimar los parámetros podría infrarrepresentar la incertidumbre generada por los datos perdidos si se usa un número pequeño de imputaciones. La comparación entre modelos mezcla distintas representaciones de entrada (componentes principales frente a variables originales). Finalmente, pese a haberse aplicado un criterio conservador (6 MAD) para el filtrado de atípicos, podrían haberse eliminado observaciones físicamente válidas.

Líneas Futuras. Como trabajo futuro se propone realizar una validación externa en catálogos como Kepler o TESS. Aumentar el número de imputaciones para capturar mejor la variabilidad debida a los datos perdidos. El uso de modelos jerárquicos con efectos por estrella o sistema. Aumentar la búsqueda óptima de hiperparámetros. Ampliación de la ingeniería de variables añadiendo rasgos específicos de tránsito.

8. Conclusiones

Mediante los datos tabulados de la misión K2, se han conseguido entrenar cuatro modelos capaces de estimar la probabilidad de que un objeto estelar sea clasificado como exoplaneta.

Se compararon los modelos mediante el mismo esquema de validación cruzada, obteniendo que el modelo con mejor desempeño fue XGBoost con un AUC de 0.88, el corte de Youden (0.82) ofreció una sensibilidad media de 0.87 y una especificidad media de 0.82.

En cuanto al poder predictivo de las variables se obtuvo que las variables estelares aportaron una mayor discriminación que las variables fotométricas (aunque hay que tener en cuenta que estos resultados pueden estar opacados por la alta correlación entre bandas fotométricas). Los coeficientes de la regresión logística fueron coherentes con las expectativas físicas, teniendo en cuenta la capacidad de la base de datos.

Los resultados muestran que el modelo XGBoost alcanza un rendimiento óptimo en validación cruzada agrupada por host_id. Las cinco variables con mayor importancia – sy_dist, st_logg, sy_pm, st_teff y st_rad – corresponden a:

- Sy_dist y sy_pm, informan sobre la distancia y movimiento propio y por ende sobre la calidad o contaminación de la señal.
- Sy_logg y st_teff, reflejan la gravedad superficial y temperatura efectiva relacionadas con el ruido y actividad estelar.
- St_rad, variable que determina la profundidad relativa del tránsito.

Estos rasgos son coherentes con la física de detección por tránsito (Brown, Latham, Everett, & Esquerdo, 2011), la consistencia entre imputaciones y pliegues indican robustez. No obstante, aún existen fuentes de incertidumbre relacionadas con el desbalanceo de clases y a la homogeneidad de parámetros dentro de un mismo sistema.

En conjunto, el trabajo cumple con el objetico de construir y evaluar un clasificador capaz de discernir entre exoplanetas y no, integrando la imputación múltiple y la validación cruzada, la validación externa, mayor número de imputaciones y gestión de la homogeneidad intrasistema permitirán refinar la precisión y aumentar la transición a otros catálogos.

9. Bibliografía

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. (2011). Kepler Input Catalog: photometric calibration and stellar classification. *The astronomical journal*. doi:10.1088/0004-6256/142/4/112

Bruno, G. (1584). De l'infinito, universo e mondi.

Bühlmann, D. J. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA: ACM.

Clarke-Pearson, D., DeLong, E., & DeLong, D. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 837-845.

Enders, C. K. (2010). Applied Missing Data Analysis. En C. K. Enders, *Applied Missing Data Analysis* (págs. 187-189). New York: Guilford.

Enders, C. K. (2010). *Applied Missing Data Analysis.* New York: Guilford.

Fischer, D. A. (2008). The Planet-Metallicity Correlation. *The Astrophysical Journal*. doi:10.1086/428383

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 3-42.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 549-553.

Howell, D. C. (2005). Median Absolute Deviation. *Encyclopedia of Statistics in Behavioral Science*.

Howell, S. B., Sobeck, C., Haas, M., Still, M., Barclay, T., Mullally, F., . . . Fortney, J. (2014). The K2 Mission: Characterization and Early Results. *Publications of the Astronomical Society of the Pacific*, 398-408. doi:10.1086/676406

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). *Journal of Statistical Software*, 1-31. doi:https://doi.org/10.18637/jss.v056.i06

Jolliffe, I. (2002). Principal Component Analysis (2nd ed.). Springer.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 141-151. doi:https://doi.org/10.1177/001316446002000116

Knox., D. (2024). Giordano Bruno. *The Stanford Encyclopedia of Philosophy*.

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

Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data.*

Mayor, M., & Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *nature*, 355-359.

Mediavilla, D. (17 de abril de 2025). Un estudio muestra "interesantes" señales en un exoplaneta que no son (de momento) vida extraterrestre. *El Pais*.

Mightell, K., & Van Cleve, J. (2020). *K2 Handbook.* Moffett Field, CA.

NASA Exoplanet Science Institute. (28 de Agosto de 2025). *Nasa Exoplanet Archive*. Obtenido de https://exoplanetarchive.ipac.caltech.edu/

Ortega, D. P. (2021). Detección de exoplanetas por el método de los tránsitos: Una simulación en Arduino. *Revista Española de Física*, 1-3.

Real Academia Española. (s. f.). Recuperado el 2 de Mayo de 2025, de https://dle.rae.es/exoplaneta

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581-592.

Rubin, D. B. (1987). *Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our View of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037/1082-989X.7.2.147

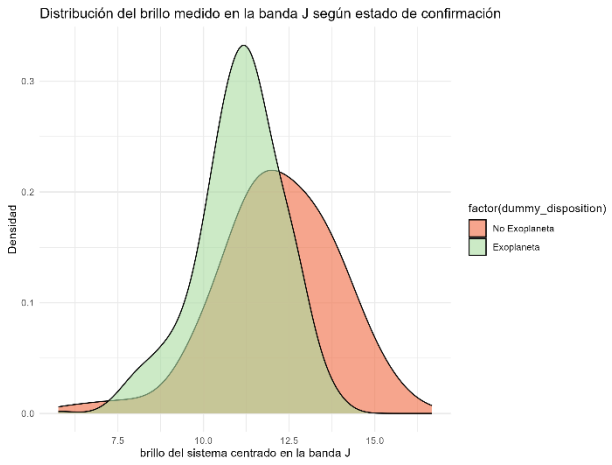
Wolszczan, A., & Frail, D. A. (1992). A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 145-147. doi:10.1038/355145a0

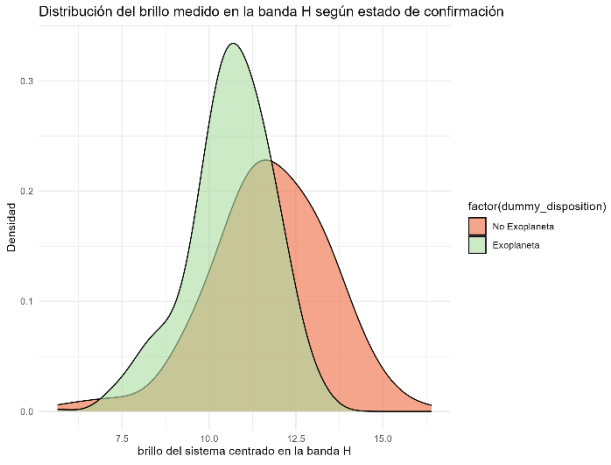
10. Anexo

A continuación, se presenta el análisis descriptivo exploratorio de las bandas restantes junto al contraste de la t de Welch.

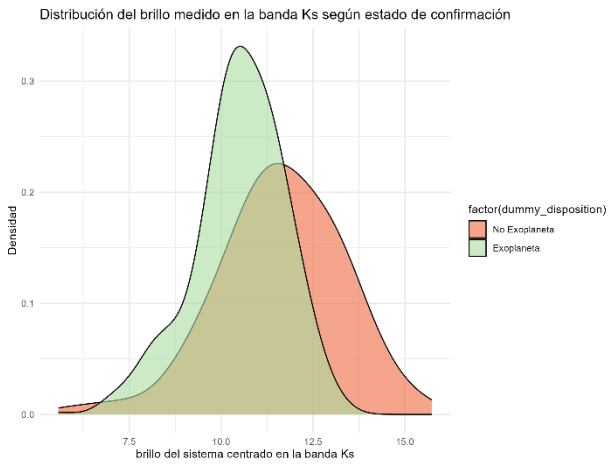
Si el p-valor obtenido es inferior a 0.05, se concluye que sí existen diferencias significativas entre las distribuciones de la banda respectiva según pertenezca al grupo de exoplaneta confirmado o no.

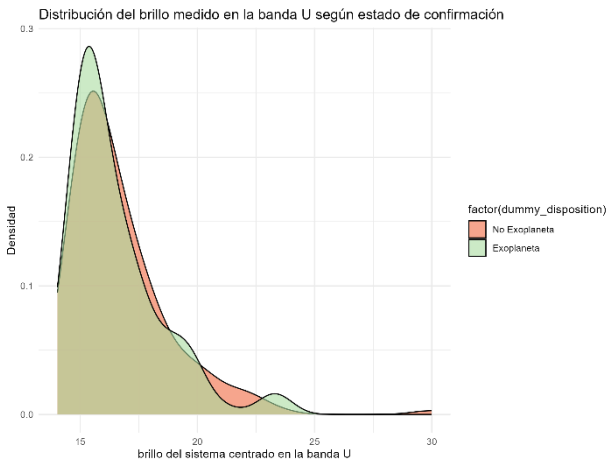
Variables sy_jmag y sy_hmag, con un estadístico t de welch de 8.57 y 9.21 respectivamente, ambos p-valores inferiores a 0.05.



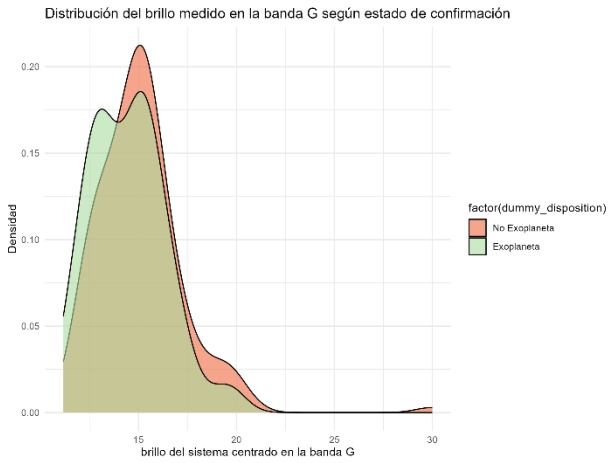


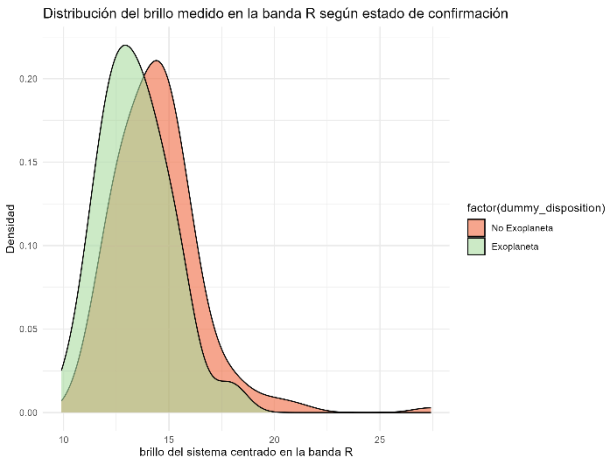
Variables sy_kmag y sy_umag, con un estadístico t de Welch de 9.19 y 0.93 respectivamente, el p-valor asociado a la variable sy_kmag es inferior a 0.05, en cambio, para el contraste para la variable sy_hmag se ha obtenido un p-valor = 0.35.



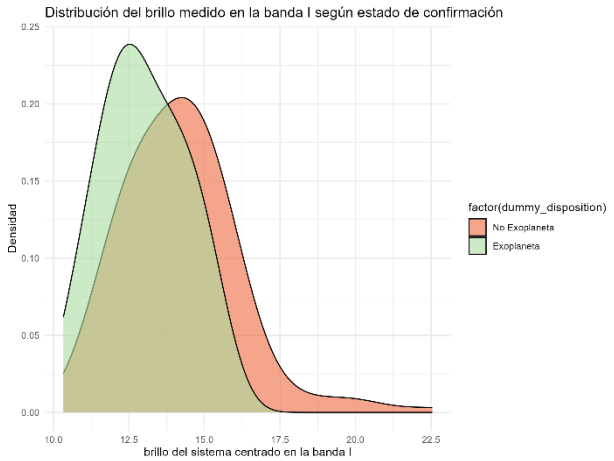


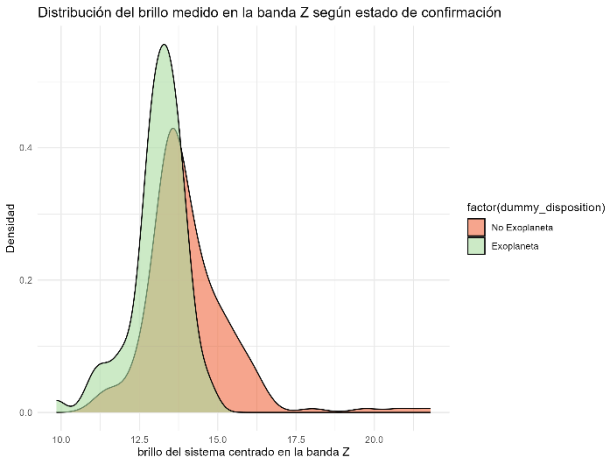
Variables sy_gmag y sy_rmag, con un estadístico t de Welch de 3.07 y 5.16 respectivamente, ambos p-valores inferiores a 0.05.



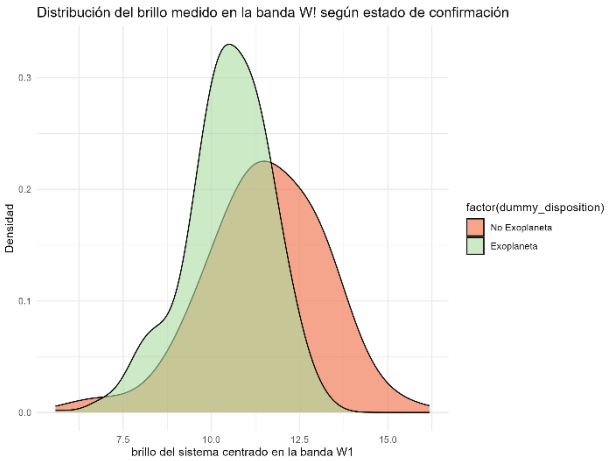


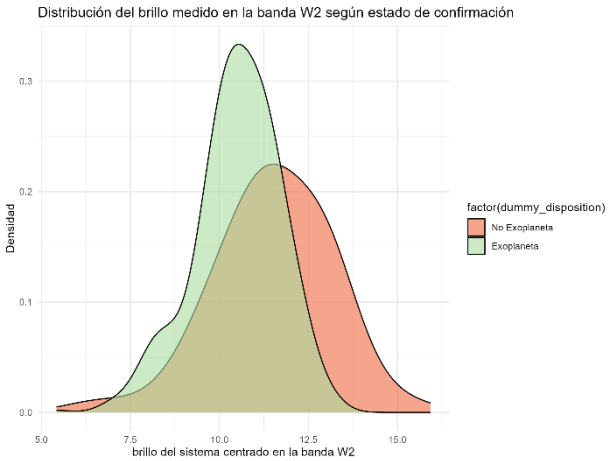
Variables sy_imag y sy_zmag, con un estadístico t de Welch de 6.69 y 7.83 respectivamente, ambos p-valores inferiores a 0.05.



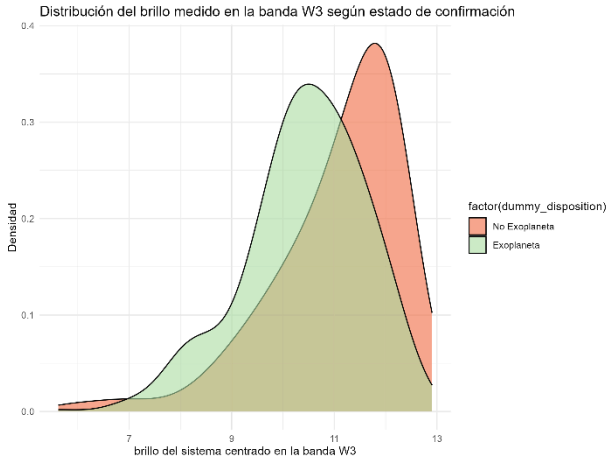


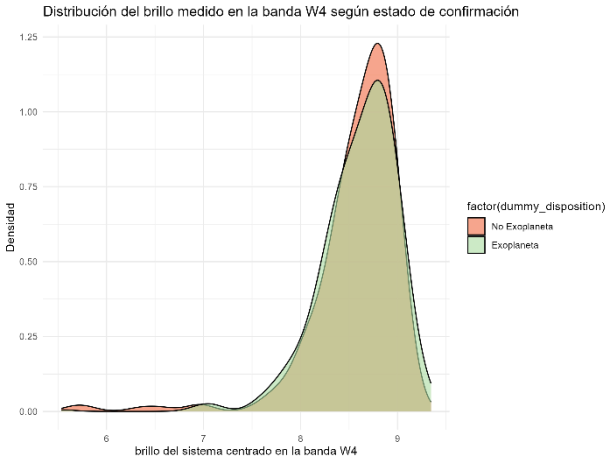
Variables sy_w1mag y sy_w2mag, con un estadístico t de Welch de 8.62 y 8.31 respectivamente, ambos p-valores inferiores a 0.05.





Variables sy_w3mag y sy_w4mag con estadísticos 6.55 y -0.86, el p-valor del contraste asociado a la variable sy_w3mag es inferior a 0.05, en cambio, en el contraste respectivo a la variable sy_w4mag se obtuvo un p-valor = 0.39





Por último, la variable sy_kepmag, con un estadístico t de Welch de 7.3 y un p-valor inferior a 0.05