A logo with blue text

Description automatically generated

Universidad Austral

Facultad de Ingeniería

Maestría en Ciencia de Datos

**Regresión Avanzada 2024**

**Trabajo Final**

**Detección de exoplanetas y sus características**

**Profesoras:**

**CHAN, Débora**

**OLIVA, Cecilia**

**Alumno:**

**NICOLAU, Jorge Enrique**

**2024**

[Resumen 3](#_Toc189694777)

[Palabras clave 3](#_Toc189694778)

[Abstract 3](#_Toc189694779)

[Keywords 3](#_Toc189694780)

[Detección de exoplanetas y sus características 4](#_Toc189694781)

[El problema 4](#_Toc189694782)

[Antecedentes 5](#_Toc189694783)

[Trabajos relacionados 6](#_Toc189694784)

[Análisis Exploratorio 8](#_Toc189694785)

[El dataset de transacciones financieras y complementarios 9](#_Toc189694786)

[Preprocesamiento 10](#_Toc189694787)

[Predicción de datos con análisis de regresión 11](#_Toc189694788)

[Regresión lineal para predicción de gastos de tarjeta de crédito 11](#_Toc189694789)

[Regresión logística para detección de eventos anómalos 11](#_Toc189694790)

[Conclusiones 12](#_Toc189694791)

[Referencias 13](#_Toc189694792)

# Resumen

Este estudio emplea **regresión lineal y logística** para analizar datos de la misión **Kepler** y predecir tanto el **radio de los exoplanetas** como su **habitabilidad**. En la regresión lineal, se modela el radio planetario (**pl\_radj**) en función de variables como la relación de radio estelar (**pl\_ratdor**), la profundidad del tránsito (**pl\_trandep**) y la amplitud de velocidad radial (**pl\_rvamp**). Los resultados indican un ajuste significativo, explicando **el 84.25% de la variabilidad** en los datos, lo que sugiere que el modelo es una herramienta útil para estimar el tamaño de los exoplanetas.

Por otro lado, la regresión logística se utiliza para definir un **Índice de Habitabilidad Planetaria (PHI)** basado en factores como temperatura de equilibrio (**pl\_eqt**), insolación (**pl\_insol**) y semieje mayor (**pl\_orbsmax**). Sin embargo, la falta de datos completos reduce la muestra a **solo 17 observaciones**, lo que genera **sobreajuste y alta incertidumbre** en las predicciones. Para mejorar la calidad del modelo, se recomienda aplicar **técnicas de imputación de datos** y explorar enfoques de **aprendizaje automático**, lo que permitiría mejorar la estabilidad y confiabilidad de los resultados.

# Palabras clave

Exoplanetas, regresión lineal, regresión logística, habitabilidad planetaria, Índice de Habitabilidad Planetaria (PHI), modelo predictivo, Kepler/K2, aprendizaje automático, valores faltantes, sobreajuste.

# Abstract

This study applies **linear and logistic regression** to analyze data from the **Kepler mission** and predict both the **radius of exoplanets** and their **habitability**. In the linear regression model, planetary radius (**pl\_radj**) is estimated based on variables such as stellar radius ratio (**pl\_ratdor**), transit depth (**pl\_trandep**), and radial velocity amplitude (**pl\_rvamp**). The results indicate a strong fit, explaining **84.25% of the variability**, suggesting that the model is a valuable tool for estimating exoplanet size.

On the other hand, logistic regression is used to define a **Planetary Habitability Index (PHI)** based on equilibrium temperature (**pl\_eqt**), insolation (**pl\_insol**), and semi-major orbital axis (**pl\_orbsmax**). However, the lack of complete data reduces the sample to **only 17 observations**, leading to **overfitting and high uncertainty** in the predictions. To improve model performance, it is recommended to apply **data imputation techniques** and explore **machine learning approaches**, which could enhance the stability and reliability of the results.

# Keywords

Exoplanets, linear regression, logistic regression, planetary habitability, Planetary Habitability Index (PHI), predictive model, Kepler/K2, machine learning, missing values, overfitting.

# Detección de exoplanetas y sus características

## El problema

Durante casi tres décadas de investigación, la NASA ha confirmado la existencia de más de 5,600 exoplanetas en 4,151 sistemas planetarios. Estos mundos, conocidos como exoplanetas o planetas extrasolares, orbitan estrellas distintas al Sol. Aunque los astrónomos los imaginaron durante milenios, no fue hasta mediados de la década de 1990 cuando comenzaron a aparecer en los registros científicos.

Aproximadamente dos tercios de los exoplanetas descubiertos hasta ahora provienen del telescopio espacial Kepler, mientras que cientos más han sido identificados gracias a la misión TESS y otros observatorios terrestres. Estas misiones han generado una enorme cantidad de datos, con cerca de 10,000 posibles exoplanetas aún pendientes de confirmación. Un número impresionante si consideramos que hasta principios de los años noventa no se conocía ninguno, pero diminuto en comparación con los cientos de miles de millones que podrían existir solo en la Vía Láctea.

Para detectar estos mundos, los astrónomos emplean diversas técnicas, muchas de ellas llevadas al límite de la tecnología disponible en los observatorios espaciales (Nardi, 2024):

### Astrometría (detección a través del movimiento estelar) (Wu, 2023)

* Velocidades radiales (detección a través del efecto Doppler) (Marín, 2018)
* Método del tránsito (detección a través de la sombra planetaria) (Alonso, 2006)
* Microlente gravitacional (detección a través de la curvatura del espacio-tiempo) (Marín, 2021)
* Observación directa (fotografía en medio del brillo estelar) (NASA, 2022)
* Detección por cómo afectan la rotación de púlsares (Wolszczan, 1992)

Los principales desafíos al encarar este tipo de análisis incluyen:

* **Multicolinealidad**: los exoplanetas presentan muchas variables orbitales, estelares y planetarias están altamente correlacionadas debido a su origen físico común (por ejemplo, la relación entre el período orbital y el semieje mayor según la Ley de Kepler).
* **Ruido y variabilidad**: las muchas variables a partir de las cuales se detectan exoplanetas suelen presentar alta variabilidad debido a la incertidumbre en las mediciones (como errores en la estimación de masas y radios planetarios) y la diversidad natural de los sistemas estelares.
* **No linealidad**: las relaciones entre propiedades estelares, orbitales y planetarias pueden ser no lineales, como en el caso de la dependencia cúbica entre el semieje mayor y el período orbital (Ley de Kepler).
* **Dimensionalidad**: la cantidad de sensores para detectar exoplanetas por diferentes metodologías pueden redundar en una gran cantidad de variables muchas de las cuales pueden ser irrelevantes o redundantes y deben descartarse para no aumentar la complejidad del modelo.

## Antecedentes

La detección de exoplanetas es un campo de investigación en constante evolución, con nuevos descubrimientos y técnicas emergentes que amplían nuestro conocimiento del universo. La ciencia de datos juega un papel fundamental en este proceso, permitiendo a los astrónomos analizar grandes volúmenes de datos y extraer información relevante sobre los exoplanetas y sus características.

Incluso **NASA** anima a la detección amateur de exoplanetas a traves de su sitio Exoplanet Watch donde ofrece una guía detallada sobre cómo analizar observaciones de tránsitos de exoplanetas para generar curvas de luz, que representan las variaciones en el brillo de una estrella cuando un planeta pasa frente a ella **(**Brachman**, 2024)**

## Trabajos relacionados

Hay varios antecedentes del uso de aprendizaje automático y métodos estadísticos para la detección de exoplanetas

En (Malik, 2021) **se** presenta una nueva técnica basada en **machine learning** para detectar exoplanetas mediante el **método de tránsito**. Se empleó la biblioteca **TSFresh** para extraer **789 características** de curvas de luz y entrenar un clasificador de **gradient boosting** con **LightGBM**.

En (Hadrien, 2025) se trata la problemática de **detección de imagen directa de exoplanetas** y cómo separar el ruido de fondo de las señales planetarias. Los métodos estadísticos recientes evitan la auto sustracción de señales de interés, a diferencia del enfoque inicial basado en **imágenes diferenciales angulares (ADI)**. Sin embargo, estos métodos pueden generar **muchos falsos positivos** si no se establecen umbrales conservadores, lo que a su vez puede hacer que se pierdan exoplanetas débiles. Este estudio extiende un marco estadístico incorporando una **regresión logística** para filtrar candidatos, utilizando características ópticas en dos longitudes de onda. Se aplica **detección de bordes y algoritmos de clustering** para procesar sub-imágenes.

En (Cardenas, 2022) se desarrolla un ensamblado **de software autónomo** para detectar tránsitos planetarios en **curvas de luz estelares**, utilizando un **clasificador de lógica difusa** y evitando la búsqueda manual de tránsitos en grandes volúmenes de datos.

En (Venkata, 2023) se presenta una técnica para detectar exoplanetas **usando el método de tránsito**. El objetivo es mejorar las técnicas tradicionales en astronomía con el uso de algoritmos de aprendizaje automático. Para ello, utilizan seis modelos diferentes de aprendizaje automático: **Random Forest, Decision Tree, Support Vector Classifier, K-Nearest Neighbor y Multi-Layer Perceptron**. Al comparar las precisiones concluye que combinando cuatro de esos modelos (Support Vector Classifier, K-Nearest Neighbor, Random Forest y Multi-Layer Perceptron) usando **bagging**, se lograba una mayor precisión.

En (Pimentel, 2024) se centra en detectar posibles planetas utilizando el método de tránsito con un enfoque novedoso al clasificar estrellas mediante un conjunto de características extraídas de series temporales en tres dominios: **temporal, estadístico y espectral**. Estas características se utilizan para entrenar y evaluar modelos con datos del **telescopio Kepler**, y los resultados superan algunos enfoques existentes. Además, el proceso de **validación cruzada** se emplea para eliminar sesgos y evaluar mejor los modelos.

inspirándose en estos antecedentes, este trabajo busca **utilizar la** **regresión parámetros faltantes de los exoplanetas** así como **predecir su habitabilidad**.

# Análisis Exploratorio

En (Tuckey, 1977) se explica que el análisis exploratorio de datos (EDA) se enfoca en descubrir patrones, anomalías y relaciones dentro de los datos, utilizando métodos visuales y estadísticos. En su base, implica abordar la incertidumbre y la diversidad de estructuras posibles en los datos, evitando asumir un único modelo probabilístico como absoluto. El EDA enfatiza la exploración sobre la inferencia estricta, buscando métodos robustos y flexibles que funcionen en diversas circunstancias, con el objetivo de proporcionar una comprensión preliminar y preparar el terreno para análisis más profundos.

## El dataset de expolanetas detectados con el telescopio espacial Kepler

El Archivo de Exoplanetas de la NASA es un catálogo astronómico en línea y un servicio de datos que recopila, correlaciona y organiza información sobre exoplanetas y sus estrellas anfitrionas. Además, proporciona herramientas para trabajar con estos datos. Este archivo está dedicado a la recopilación y difusión de conjuntos de datos públicos claves utilizados en la búsqueda y caracterización de planetas extrasolares y sus estrellas (IPAC, 2021).

Al momento de la extracción (14 de enero de 2025) el dataset de exoplanetas contiene información sobre 28.217 exoplanetas (entre confirmados, propuestos y en estudio) y sus características. La base de datos de IPAC contiene información sobre los exoplanetas detectados por la misión Kepler y otros telescopios, pero para este trabajo se utilizará solo la información de los exoplanetas detectados por la misión Kepler (Cermak, 2024).

## Preprocesamiento

Se cargarán los datos, se filtrarán las filas de los exoplanetas confirmados, las columnas relevantes, se eliminarán los valores faltantes y se realizará un análisis exploratorio de los datos para luego realizar la regresión lineal y logística. Los datos iniciales están en **keplerexoplanets.csv** en el repositorio de este trabajo (Nicolau, 2025). El mismo se obtuvo de la página del IPAC del Caltech (IPAC, 2025).

El dataset contiene información sobre los sistemas exoplanetarios detectados por la misión Kepler y el diccionario de datos se encuentra en línea (IPAC, 2024) explicando en más profundidad la semántica de las variables.

Para el análisis se considerarán solo los **exoplanetas confirmados y se eliminarán las columnas no relevantes para el análisis**.

Para reducir la dimensionalidad del problema, a continuación se **eliminan las columnas con referencias a sitios web no relevantes para el análisis**. También se eliminan los **datos de referencia de los planetas, estrellas y sistemas no relevantes** para el análisis tales como referencias a catálogos externos de estrellas y sistemas, **información de publicación del descubrimiento, información de detección, información de fotometría**, banderas de organización interna del dataset, información del sistema planetario y detalles técnicos de la detección.

Para reducir la cantidad de registros del dataset y como se busca predecir la habitabilidad de los exoplanetas en sistemas similares al Solar, **se filtran los sistemas con estrellas de tamaño similar al Sol (0.9 a 1.1 radios solares)**. Además se considerarán para la muestra **solo los exoplanetas publicados en la literatura científica**, por lo que se filtra por “Published Confirmed” en la columna **soltype**.

El dataset resultante del preprocesado **contiene 2547 registros y 68 columnas** y tiene estas columnas:

* **Posición y Coordenadas.** La ubicación de la estrella en el cielo se expresa en diferentes sistemas de coordenadas: la **ascensión recta** (**ra, rastr**) y la **declinación** (**dec, decstr**) en coordenadas ecuatoriales, además de las **coordenadas galácticas** (**glat, glon**) y **eclípticas** (**elat, elon**). También se incluye la **distancia al sistema** (**sy\_dist**), su **paralaje** (**sy\_plx**) y su **movimiento propio** (**sy\_pm, sy\_pmra, sy\_pmdec**).
* **Identificación del Sistema Estelar.** El sistema estelar se puede identificar mediante diversos nombres provenientes de catálogos astronómicos, como el **Henry Draper** (**hd\_name**), **Hipparcos** (**hip\_name**) y el **nombre comúnmente usado en la literatura** **científica** (**hostname**).
* **Propiedades del Planeta.** Los exoplanetas tienen diversas características físicas y orbitales. Su **masa** se proporciona en diferentes unidades, como masas terrestres (**pl\_masse, pl\_bmasse, pl\_cmasse**) y jovianas (**pl\_massj, pl\_bmassj, pl\_cmassj**). Se incluyen valores de masa obtenidos por diferentes métodos, como la **mínima proyectada** (**pl\_cmasse, pl\_cmassj**) y la determinada por **velocidad radial** (**pl\_msinie, pl\_msinij**). También se reportan la **densidad** (**pl\_dens**), el **radio** (**pl\_rade, pl\_radj**), y su relación con el radio estelar (**pl\_ratdor, pl\_ratror**). El comportamiento orbital del planeta se describe con la **excentricidad** (**pl\_orbeccen**), la **inclinación orbital** (**pl\_orbincl**), el **argumento del periastro** (**pl\_orblper**), el **semieje mayor** (**pl\_orbsmax**), el **período orbital** (**pl\_orbper**), y el **momento del periastro** (**pl\_orbtper**). También se incluye información sobre su **insolación** (**pl\_insol**), la **temperatura de equilibrio** (**pl\_eqt**), y parámetros relacionados con su tránsito, como la **profundidad del tránsito** (**pl\_trandep**), su **duración** (**pl\_trandur**) y la **amplitud de velocidad radial** (**pl\_rvamp**).
* **Propiedades de la Estrella.** Las estrellas que albergan exoplanetas están descritas en términos de su **edad** (**st\_age**), **densidad** (**st\_dens**), **gravedad superficial** (**st\_logg**), **luminosidad** (**st\_lum**), **masa** (**st\_mass**), y **radio** (**st\_rad**). También se incluyen medidas de su **metalicidad** (**st\_met, st\_metratio**), su **temperatura efectiva** (**st\_teff**), y su **tipo espectral** (**st\_spectype**). La información sobre su **movimiento y rotación** se expresa mediante la **velocidad radial** (**st\_radv**), la **velocidad de rotación** (**st\_vsin**), y el **período de rotación** (**st\_rotp**).
* **Información del Sistema Planetario.** El número de cuerpos en el sistema se indica con **sy\_pnum** para planetas y **sy\_snum** para estrellas. También se incluye el **método de descubrimiento** (**discoverymethod**) y un **flag de tránsito de tiempo variante** (**ttv\_flag**). En sistemas binarios, el parámetro **cb\_flag** señala si el planeta orbita un sistema estelar doble.

## Revisión de variables numéricas

| **Variable** | **Media** | **Des.Est.** | **Mín.** | **P25** | **Mediana** | **P75** | **Máx.** | **Hist.** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **cb\_flag** | 0.0024 | 0.0485 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | ▇▁▁▁▁ |
| **dec** | 44.2366 | 3.6695 | 36.5773 | 41.3738 | 44.3155 | 47.1127 | 52.1491 | ▅▇▇▇▃ |
| **elat** | 64.7746 | 3.6360 | 57.6374 | 61.8480 | 64.7690 | 67.5246 | 72.4913 | ▅▇▇▆▃ |
| **elon** | 307.9950 | 8.0410 | 288.7515 | 302.3200 | 308.2267 | 314.3187 | 323.9048 | ▂▅▇▇▃ |
| **glat** | 12.9778 | 3.4193 | 5.9607 | 10.4051 | 12.6586 | 15.7959 | 21.1430 | ▃▇▇▅▂ |
| **glon** | 76.3718 | 3.6938 | 68.2330 | 73.3769 | 76.4011 | 79.3326 | 84.3822 | ▃▇▇▇▃ |
| **pl\_bmasse** | 328.4638 | 1965.7298 | 0.7600 | 6.0133 | 18.4000 | 95.1000 | 25426.4000 | ▇▁▁▁▁ |
| **pl\_bmassj** | 1.0335 | 6.1849 | 0.0024 | 0.0190 | 0.0580 | 0.2992 | 80.0000 | ▇▁▁▁▁ |
| **pl\_dens** | 4.2419 | 8.2500 | 0.0300 | 0.6950 | 2.1300 | 5.5250 | 77.7000 | ▇▁▁▁▁ |
| **pl\_eqt** | 938.0333 | 564.4973 | 251.0000 | 438.7500 | 861.5000 | 1146.5000 | 2188.0000 | ▇▆▃▂▂ |
| **pl\_imppar** | 0.3940 | 0.2883 | 0.0000 | 0.1500 | 0.3300 | 0.6200 | 1.4830 | ▇▅▃▁▁ |
| **pl\_insol** | 299.3455 | 598.0846 | 0.5900 | 26.9900 | 96.9450 | 304.9000 | 4849.2600 | ▇▁▁▁▁ |
| **pl\_masse** | 328.4638 | 1965.7298 | 0.7600 | 6.0133 | 18.4000 | 95.1000 | 25426.4000 | ▇▁▁▁▁ |
| **pl\_massj** | 1.0335 | 6.1849 | 0.0024 | 0.0190 | 0.0580 | 0.2992 | 80.0000 | ▇▁▁▁▁ |
| **pl\_msinie** | 47.8000 | NA | 47.8000 | 47.8000 | 47.8000 | 47.8000 | 47.8000 | ▁▁▇▁▁ |
| **pl\_msinij** | 0.1500 | NA | 0.1500 | 0.1500 | 0.1500 | 0.1500 | 0.1500 | ▁▁▇▁▁ |
| **pl\_orbeccen** | 0.1177 | 0.1778 | 0.0000 | 0.0105 | 0.0420 | 0.1350 | 0.8380 | ▇▁▁▁▁ |
| **pl\_orbincl** | 88.6474 | 1.7458 | 82.2140 | 87.8060 | 89.1855 | 89.7905 | 93.1500 | ▁▁▃▇▁ |
| **pl\_orblper** | 146.9116 | 142.8866 | -163.0000 | 49.2755 | 154.7000 | 261.6000 | 357.0300 | ▂▅▆▅▇ |
| **pl\_orbper** | 31.4997 | 62.6948 | 0.5383 | 5.6992 | 13.0314 | 33.6013 | 1322.3000 | ▇▁▁▁▁ |
| **pl\_orbsmax** | 0.1834 | 0.2236 | 0.0168 | 0.0655 | 0.1091 | 0.2170 | 2.4200 | ▇▁▁▁▁ |
| **pl\_orbtper** | 2455003.7584 | 106.5838 | 2454935.8000 | 2454950.2605 | 2454958.2168 | 2455011.7148 | 2455162.8000 | ▇▁▁▁▂ |
| **pl\_projobliq** | -35.1429 | 60.4109 | -135.0000 | -61.5000 | 0.0000 | 4.0000 | 4.0000 | ▃▁▁▁▇ |
| **pl\_rade** | 2.9323 | 2.4308 | 0.4000 | 1.6230 | 2.3700 | 3.0240 | 30.8000 | ▇▁▁▁▁ |
| **pl\_radj** | 0.2616 | 0.2169 | 0.0360 | 0.1450 | 0.2110 | 0.2700 | 2.7480 | ▇▁▁▁▁ |
| **pl\_ratdor** | 60.9640 | 83.9114 | 3.1000 | 11.5600 | 30.3000 | 78.3000 | 576.7000 | ▇▁▁▁▁ |
| **pl\_ratror** | 0.0288 | 0.0263 | 0.0056 | 0.0147 | 0.0213 | 0.0286 | 0.2873 | ▇▁▁▁▁ |
| **pl\_rvamp** | 47.4204 | 88.2590 | 0.2800 | 2.1000 | 3.7200 | 71.9000 | 419.5000 | ▇▂▁▁▁ |
| **pl\_trandep** | 0.1813 | 0.3624 | 0.0060 | 0.0481 | 0.0705 | 0.1063 | 2.2620 | ▇▁▁▁▁ |
| **pl\_trandur** | 4.3811 | 2.2576 | 0.8244 | 2.8100 | 3.8212 | 5.4291 | 18.8860 | ▇▅▁▁▁ |
| **pl\_tranmid** | 2455003.7508 | 121.7668 | 2454832.9010 | 2454967.1167 | 2454973.9640 | 2455004.6853 | 2457959.9661 | ▇▁▁▁▁ |
| **ra** | 291.3112 | 4.6829 | 280.2066 | 287.6980 | 291.4314 | 295.0027 | 301.5430 | ▂▆▇▇▂ |
| **st\_age** | 4.5799 | 1.7720 | 0.1050 | 3.8000 | 4.2700 | 4.7900 | 11.9000 | ▁▇▂▁▁ |
| **st\_dens** | 1.6002 | 1.3618 | 0.0044 | 1.2072 | 1.4497 | 1.6696 | 19.9316 | ▇▁▁▁▁ |
| **st\_logg** | 4.4442 | 0.0556 | 4.2100 | 4.4100 | 4.4500 | 4.4800 | 5.0000 | ▁▇▁▁▁ |
| **st\_lum** | -0.0616 | 0.1126 | -0.4090 | -0.1360 | -0.0630 | 0.0280 | 0.2200 | ▁▃▇▆▂ |
| **st\_mass** | 0.9901 | 0.0642 | 0.6900 | 0.9500 | 0.9900 | 1.0300 | 1.2500 | ▁▁▇▃▁ |
| **st\_met** | 0.0199 | 0.1570 | -0.8160 | -0.0400 | 0.0200 | 0.1000 | 0.4800 | ▁▁▃▇▁ |
| **st\_rad** | 0.9922 | 0.0619 | 0.9000 | 0.9400 | 0.9900 | 1.0500 | 1.1000 | ▇▅▅▅▅ |
| **st\_radv** | -34.3600 | 29.8608 | -98.9300 | -57.1600 | -24.7600 | -20.9300 | 9.9600 | ▂▃▂▇▃ |
| **st\_rotp** | 12.2225 | 5.6116 | 4.6900 | 10.1775 | 11.8900 | 14.3225 | 22.0500 | ▃▁▇▂▂ |
| **st\_teff** | 5684.9878 | 236.0809 | 4388.3900 | 5549.5000 | 5688.0000 | 5833.5000 | 6484.0000 | ▁▁▆▇▁ |
| **st\_vsin** | 2.2671 | 1.9863 | 0.3000 | 0.5000 | 2.0000 | 3.0000 | 10.4000 | ▇▃▁▁▁ |
| **sy\_dist** | 875.6348 | 389.0261 | 68.1730 | 625.3390 | 860.3910 | 1101.1000 | 2879.8300 | ▃▇▂▁▁ |
| **sy\_plx** | 1.4880 | 1.2326 | 0.3269 | 0.8774 | 1.1298 | 1.5694 | 14.6396 | ▇▁▁▁▁ |
| **sy\_pm** | 10.9292 | 9.0598 | 0.1586 | 4.5898 | 8.4847 | 14.5826 | 77.6183 | ▇▂▁▁▁ |
| **sy\_pmdec** | -3.2419 | 11.7992 | -66.6869 | -9.1002 | -3.1269 | 2.5670 | 48.3025 | ▁▁▇▃▁ |
| **sy\_pmra** | -0.3364 | 7.1920 | -49.3144 | -3.5431 | -0.4287 | 3.0154 | 32.8344 | ▁▁▇▇▁ |
| **sy\_pnum** | 2.0628 | 1.2462 | 1.0000 | 1.0000 | 2.0000 | 3.0000 | 6.0000 | ▇▂▁▁▁ |
| **sy\_snum** | 1.0271 | 0.1624 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 2.0000 | ▇▁▁▁▁ |
| **ttv\_flag** | 0.1496 | 0.3567 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | ▇▁▁▁▂ |

El análisis revela que las variables binarias **cb\_flag** y **ttv\_flag** están fuertemente desbalanceadas, con valores mayoritariamente en 0, lo que indica que la mayoría de los sistemas no son circumbinarios ni presentan tránsitos de tiempo variable. Además, la variable **sy\_snum** sugiere que la mayoría de los sistemas tienen una sola estrella.

En cuanto a las coordenadas celestes, las distribuciones de **dec** y **ra** son relativamente centradas, mientras que **elon** presenta un rango amplio (288°-324°), abarcando las constelaciones de Capricornio, Acuario y Piscis. La latitud eclíptica **elat** es alta (57°-72°), lo que indica que los objetos están lejos del plano de la eclíptica y se concentran en la región del Polo Norte Eclíptico, cerca de la constelación del Dragón.

Las variables de masa planetaria (**pl\_bmasse, pl\_masse, pl\_bmassj, pl\_massj**) muestran distribuciones sesgadas a la derecha con valores extremos, lo que indica la presencia de algunos planetas muy masivos. De manera similar, las distribuciones de radio (**pl\_rade, pl\_radj**) están sesgadas, con la mayoría de los planetas siendo más grandes que la Tierra (mediana de **pl\_rade** ≈ 2.37). La excentricidad orbital (**pl\_orbeccen**) es predominantemente baja, sugiriendo órbitas casi circulares, mientras que la inclinación orbital (**pl\_orbincl**) está centrada en 90°, lo que es consistente con la detección por el método de tránsito. El período orbital (**pl\_orbper**) tiene una gran variabilidad, con una mediana de 13 días, pero valores extremos de hasta 1,322 días.

La temperatura de equilibrio planetaria (**pl\_eqt**) muestra una amplia distribución (251 K - 2,188 K), lo que indica la presencia de tanto planetas fríos como extremadamente calientes. La densidad (**pl\_dens**) es variable, con una mediana de 2.13 g/cm³, reflejando la diversidad en la composición planetaria.

Respecto a las estrellas, la mayoría tienen masas y radios similares al Sol (**st\_mass, st\_rad ≈ 1.0**), con temperaturas efectivas (**st\_teff**) cercanas a 5,685 K, típicas de estrellas tipo G. La gravedad superficial (**st\_logg**) está centrada en 4.45, lo que es característico de estrellas en la secuencia principal. La luminosidad (**st\_lum**) también es similar a la del Sol en su mayoría.

El análisis de distancias muestra que los sistemas están relativamente lejos, con una mediana de 860 parsecs (≈2,804 años luz). El movimiento propio (**sy\_pm, sy\_pmra, sy\_pmdec**) está mayormente centrado en 0, aunque algunas estrellas tienen movimientos más rápidos. El número de planetas por sistema (**sy\_pnum**) tiene una mediana de 2, con un máximo de 6 planetas detectados.

Finalmente, los parámetros de tránsito (**pl\_trandep, pl\_trandur**) muestran distribuciones sesgadas, indicando que la mayoría de los tránsitos son poco profundos y de corta duración. La oblicuidad (**pl\_projobliq**) es mayormente 0°, pero hay valores extremos de hasta -135°. En general, las distribuciones presentan sesgos y valores extremos, reflejando tanto la diversidad de los exoplanetas como la presencia de valores atípicos. La mayoría de los exoplanetas en el conjunto de datos son relativamente pequeños, aunque existen algunos extremadamente grandes, y las estrellas tienden a ser similares al Sol con variaciones en edad y metalicidad, en parte debido al filtro aplicado en el preprocesamiento.

## Detención de valores faltantes

Se analiza la cantidad de valores faltantes en el dataset para identificar posibles problemas de calidad de datos.

A graph with blue lines

Description automatically generated with medium confidence

A graph with numbers and a bar

Description automatically generated

Como parte del proceso, se analizarán las **variables conocidas de los exoplanetas** para identificar aquellas que puedan actuar como **predictoras** de las variables con valores faltantes. Por ejemplo, en el caso del **radio del exoplaneta** (representado por **pl\_radj** en radios jovianos y **pl\_rade** en radios terrestres), que presenta **886 valores faltantes**, es posible estimarlo utilizando otras variables disponibles en el dataset. Algunas de estas incluyen el **período orbital en días** (**pl\_orbper**, con **450 valores faltantes**), el **tiempo de conjunción** (**pl\_tranmid**, con **675 valores faltantes**) y el **logaritmo en base 10 de la gravedad superficial de la estrella del sistema planetario** (**st\_logg**, con **554 valores faltantes**).

## Buscando correlaciones

Se analiza la correlación entre las variables numéricas del dataset para identificar posibles relaciones entre ellas utilizando los métodos de Pearson (detección de correlaciones lineales). A través de esta matriz, se pueden detectar tanto **correlaciones positivas** (cuando un aumento en una variable está asociado con un aumento en otra) como **correlaciones negativas** (cuando un aumento en una variable está relacionado con una disminución en otra). Estas relaciones pueden ser útiles para entender patrones en los datos y seleccionar las mejores variables predictoras para modelos de estimación.

A screen shot of a computer

Description automatically generated A close up of a data

Description automatically generated

A diagram of a number

Description automatically generated with medium confidence

Un ejemplo de estas correlaciones se observa en la relación entre el **cociente del semieje orbital mayor y el período orbital** (**pl\_ratdor**), el cual muestra una **correlación positiva** con el **período orbital** (**pl\_orbper**). Esto sugiere que a medida que el período orbital aumenta, también lo hace este cociente. En contraste, **pl\_ratdor** presenta una **correlación negativa** con la **temperatura de equilibrio del exoplaneta** (**pl\_eqt**), lo que indica que los exoplanetas con períodos orbitales largos tienden a tener temperaturas más bajas, lo cual es esperable, ya que suelen estar más alejados de su estrella anfitriona.

En cuanto a la variable que se seleccionó como **variable objetivo** en el análisis previo, el **radio del exoplaneta** (**pl\_radj**), se observan relaciones significativas con otras variables. Específicamente, tiene una **correlación positiva** con la **profundidad del tránsito** (**pl\_trandep**), lo que indica que los exoplanetas más grandes generan una mayor disminución en el brillo de su estrella al pasar frente a ella. Por otro lado, **pl\_radj** muestra una **correlación negativa** con la **gravedad superficial de la estrella anfitriona** (**st\_logg**), lo que sugiere que los exoplanetas más grandes tienden a encontrarse en sistemas con estrellas de menor gravedad superficial, posiblemente porque estos sistemas albergan planetas de mayor tamaño.

Más allá de la relación lineal evidente entre las medidas del **radio del exoplaneta en diferentes unidades** —es decir, **radios jovianos** y **radios terrestres**—, el análisis de correlación revela la presencia de otras relaciones lineales significativas entre el radio del exoplaneta y otras variables del conjunto de datos.

La primera variable con una fuerte correlación es el **cociente del radio del planeta sobre el radio de su estrella anfitriona** (**pl\_ratror**). La segunda variable con alta correlación es la **profundidad del tránsito** (**pl\_trandep**). La tercera variable significativa es la **amplitud de la velocidad radial del planeta** (**pl\_rvamp**).

Con base en estos resultados, se seleccionan estas tres variables como **variables predictoras** para el radio del exoplaneta, ya que presentan una **alta correlación lineal**, positiva o negativa, con la variable faltante.

## Detección de valores atípicos (outliers) en las predictoras

Para las variables predictoras seleccionadas (**pl\_ratdor**,**pl\_trandep**,**pl\_rvamp**), se detectan valores atípicos en el dataset que puedan afectar el análisis y la predicción de la variable faltante (**pl\_radj**).

A graph with numbers and lines

Description automatically generated

El **boxplot** revela la presencia de **valores atípicos** en las variables predictoras seleccionadas, lo que puede influir significativamente en el análisis y en la precisión de la estimación de la variable faltante. Estos **outliers** pueden sesgar los resultados y reducir la eficacia de los modelos de predicción, especialmente si se utilizan técnicas sensibles a valores extremos, como la regresión lineal estándar.

Para abordar este problema, es recomendable emplear **métodos de regresión robustos**, diseñados para minimizar el impacto de valores atípicos en el ajuste del modelo.

# Análisis de regresión

Para explorar la relación entre las características conocidas de los exoplanetas, se lleva a cabo un **análisis de regresión lineal**. El objetivo es predecir el **radio del exoplaneta (pl\_radj)** utilizando tres variables que podrían estar estrechamente relacionadas con su tamaño:

* **pl\_ratdor**: el cociente entre el radio del exoplaneta y el radio de su estrella, que proporciona una medida relativa del tamaño del planeta.
* **pl\_trandep**: la profundidad del tránsito, que indica cuánto disminuye el brillo de la estrella cuando el exoplaneta pasa frente a ella, lo que está directamente relacionado con su radio.
* **pl\_rvamp**: la amplitud de la velocidad radial del planeta, que refleja su efecto gravitacional sobre la estrella y puede aportar información sobre su masa y, en combinación con otras variables, su tamaño.

A través de este análisis, se busca comprender mejor cómo estas variables influyen en el tamaño de los exoplanetas y determinar cuál de ellas tiene un mayor impacto en la predicción de **pl\_radj**.

## Regresión lineal univariada

Se explorar un análisis de regresión lineal univariada con varias variables independientes evaluando la capacidad de predicción de cada modelo planteado.

**Regresión lineal: radio del exoplaneta según ratio del radio sobre el radio estelar (pl\_ratdor)**

##   
## Call:  
## lm(formula = pl\_radj ~ pl\_ratdor, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3749 -0.2695 -0.2125 0.3126 1.2264   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4086542 0.0535610 7.630 0.0000000000489 \*\*\*  
## pl\_ratdor 0.0007341 0.0005069 1.448 0.152   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3872 on 78 degrees of freedom  
## (2467 observations deleted due to missingness)  
## Multiple R-squared: 0.02618, Adjusted R-squared: 0.0137   
## F-statistic: 2.097 on 1 and 78 DF, p-value: 0.1516

A graph with a red line and blue dots

Description automatically generated

El análisis de regresión indica que **pl\_ratdor** no es un buen predictor del radio del exoplaneta (**pl\_radj**). El coeficiente de regresión es muy pequeño (0.0007341), y el valor p (0.152) es mayor que 0.05, lo que sugiere que no hay suficiente evidencia estadística para afirmar una relación significativa entre ambas variables.

Las estadísticas de ajuste refuerzan esta conclusión. El **Error Estándar Residual** (0.3872) indica una gran desviación entre los valores predichos y reales. El **R² múltiple** (0.02618) muestra que solo el 2.6% de la variabilidad en **pl\_radj** es explicada por **pl\_ratdor**, y el **R² ajustado** (0.0137) confirma la baja capacidad predictiva del modelo. Además, el **F-statistic** (2.097) y su valor p (0.1516) indican que el modelo completo no es estadísticamente significativo, lo que confirma que **pl\_ratdor** no contribuye de manera relevante a la estimación del radio del exoplaneta.

A graph with blue dots

Description automatically generated

El análisis de los residuos muestra que estos se concentran en los valores ajustados más bajos (0.4-0.5) y tienden a dispersarse a medida que los valores ajustados aumentan. No se observa un patrón curvo, lo que sugiere una relación aproximadamente lineal entre **pl\_ratdor** y **pl\_radj**.

Dado el bajo **R² (~2.6%)**, era esperable encontrar alta dispersión en los residuos, lo que confirma que **pl\_ratdor** no es un buen predictor de **pl\_radj**. La presencia de valores atípicos sugiere que existen otras variables más relevantes que explican mejor la variabilidad en el radio del exoplaneta.

A graph with red dots

Description automatically generated

El análisis de normalidad de los residuos muestra desviaciones en los cuantiles extremos, donde los puntos se alejan de la línea teórica, indicando la presencia de colas más pesadas de lo esperado en una distribución normal. Esto sugiere la existencia de valores atípicos o que los residuos no siguen completamente una distribución normal. Dado que la regresión lineal asume normalidad en los residuos para garantizar la validez de los valores p e intervalos de confianza, estas desviaciones plantean dudas sobre el cumplimiento de esta suposición.

Para mitigar este problema, se podrían aplicar transformaciones en la variable dependiente (**pl\_radj**), como logaritmos o raíces cuadradas, para mejorar la normalidad de los residuos.

##   
## Shapiro-Wilk normality test  
##   
## data: res\_pl\_ratdor  
## W = 0.81534, p-value = 0.00000001313

La prueba de **Shapiro-Wilk** se utilizó para evaluar si los residuos del modelo siguen una distribución normal, una suposición clave en la regresión lineal. Los resultados muestran un **estadístico W de 0.81534**, lo que indica una desviación considerable de la normalidad, ya que valores cercanos a 1 sugieren una distribución normal. Además, el **p-valor obtenido (≈ 0.00000001313)** es significativamente menor a 0.05, lo que proporciona evidencia sólida para rechazar la hipótesis nula.

##   
## studentized Breusch-Pagan test  
##   
## data: model\_pl\_ratdor  
## BP = 1.1306, df = 1, p-value = 0.2876

El **test de Breusch-Pagan** se utilizó para evaluar si los residuos del modelo presentan homocedasticidad (varianza constante).El **estadístico BP fue 1.1306**, un valor relativamente bajo que no indica una fuerte señal de heterocedasticidad. Esto sugiere que **no se detecta heterocedasticidad significativa en los residuos**, lo que implica que la varianza de los errores es relativamente constante en el modelo.

**Regresión lineal: radio del exoplaneta según la profundidad del tránsito (pl\_trandep)**

## Call:  
## lm(formula = pl\_radj ~ pl\_trandep, data = kepler\_data\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.203530 -0.056704 0.002386 0.056979 0.150580   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.22308 0.01590 14.03 <0.0000000000000002 \*\*\*  
## pl\_trandep 0.61058 0.02434 25.08 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08315 on 40 degrees of freedom  
## (2505 observations deleted due to missingness)  
## Multiple R-squared: 0.9402, Adjusted R-squared: 0.9387   
## F-statistic: 629.1 on 1 and 40 DF, p-value: < 0.00000000000000022

A graph with a red line and blue dots

Description automatically generated

El modelo presenta un **error estándar de los residuos de 0.08315**, indicando una baja variabilidad en las predicciones y un ajuste preciso. El **coeficiente de determinación R² es del 94.02%**, lo que significa que el modelo explica casi toda la variabilidad de **pl\_radj**. Incluso el **R² ajustado (93.87%)** sigue siendo muy alto, confirmando la solidez del ajuste.

La **prueba de significancia global**, evaluada con el **estadístico F (629.1)** y un **p-valor extremadamente bajo**, refuerza la validez del modelo y su capacidad para explicar la relación entre las variables. Comparado con modelos anteriores (**pl\_radj ~ pl\_ratdor** y **pl\_radj ~ log(pl\_ratdor)**), este nuevo modelo muestra una mejora drástica en el ajuste. La relación entre **pl\_trandep** y **pl\_radj** es fuerte y estadísticamente significativa, lo que indica que **pl\_trandep es un excelente predictor del radio del exoplaneta**.

# Conclusiones

El disertante expuso acerca del estudio realizado sobre la interacción entre la inteligencia artificial y diferentes modelos predictivos en el rubro de la agricultura. El objetivo del trabajo exhibido, desde un punto de vista técnico, es encontrar la mejor combinación de *clústeres* para, de acuerdo con el potencial rendimiento del suelo, colocar una mayor o menor cantidad de semillas. El objetivo real es maximizar el margen bruto a largo plazo y minimizar el riesgo. Asimismo, es importante destacar que, según lo explicado por el expositor, el mejor rendimiento sistemático en un caso real lo obtuvo a través de una red neuronal.

# Referencias

Alonso Sobrino, Roi. (2006) “Detección y Caracterización de Exoplanetas Mediante El Método de Los Tránsitos.” Instituto de Astrofísica de Canarias (IAC). <https://www.iac.es/es/ciencia-y-tecnologia/publicaciones/deteccion-y-caracterizacion-de-exoplanetas-mediante-el-metodo-de-los-transitos>.

Brachman, R. Z. (2024) “How to Analyze Your Data | How to Get Started – Exoplanet Exploration: Planets Beyond Our Solar System.” Exoplanet Exploration: Planets Beyond Our Solar System. <https://exoplanets.nasa.gov/exoplanet-watch/how-to-contribute/how-to-analyze-your-data/>.

Cardenas, Christian & Lozano, David & Marquez, Cristian & Torres, Edilberto & Delgado-Correal, Camilo. (2022). Optimización de un sistema difuso para la detección automática de tránsitos planetarios en curvas de luz de estrellas individuales. Ciencia en Desarrollo. 1. 19-35. 10. <http://dx.doi.org/10.19053/01217488.v1.n2E.2022.15136>.

Cermak, A. & Cermak, A. (2024) “Kepler / K2 - NASA Science.” NASA Science. <https://science.nasa.gov/mission/kepler/>.

Hadrien Cambazard, Nicolas Catusse, Antoine Chomez, Anne-Marie Lagrange (2025). Logistic regression to boost exoplanet detection performances, Monthly Notices of the Royal Astronomical Society, Volume 536, Issue 2, Pages 1610–1624, <https://doi.org/10.1093/mnras/stae2657>

IPAC. (2025) “NASA Exoplanet Archive.” NASA Exoplanet Science Institute. <https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS>.

———. (2021) “NASA Exoplanet Archive Overview and Holdings.” <https://exoplanetarchive.ipac.caltech.edu/docs/intro.html>.

———. (s.f.) “Our Mission.” <https://www.ipac.caltech.edu/page/mission>.

———. (2024) “Planetary Systems and Planetary Systems Composite Parameters Data Column Definitions.” <https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html>.

Kane, Stephen R. & Gelino, Dawn M. (2012) “The Habitable Zone and Extreme Planetary Orbits.” <https://doi.org/10.1089/ast.2011.0798>.

Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V., & Deshpande, R. (2013) “Habitable Zones Around Main-Sequence Stars: New Estimates.” <https://doi.org/10.1088/0004-637x/765/2/131>.

Malik, A., Moster, B. P., & Obermeier, C. (2021). Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society*. <https://doi.org/10.1093/mnras/stab3692>

Marín, Daniel. (2011) “Detectando Planetas Desde El Espacio Gracias a Einstein - Eureka.” Eureka. <https://danielmarin.naukas.com/2011/05/23/detectando-planetas-desde-el-espacio-gracias-a-einstein/>.

———. (2018) “EarthFinder: Un Telescopio Espacial Para Buscar Exotierras Por El Método de La Velocidad Radial - Eureka.” <https://danielmarin.naukas.com/2018/03/14/earthfinder/>.

Méndez, Abel, Rivera-Valentin, Edgard, Schulze-Makuch, Dirk, Filiberto, Justin, Ramirez, Ramses, Wood, Tana, Dávila, Alfonso, McKay, Chris, Ceballos, Kevin, Jusino-Maldonado, Marcos, Torres-Santiago, Nicole, Gomez, Guillermo Nery, Heller, René, Byrne, Paul, Malaska, Michael, Nathan, Erica, Simões, Marta, Antunes, André, Martínez-Frías, Jesús, & Haqq-Misra, Jacob. (2021) *“Habitability Models for Astrobiology.”* <https://doi.org/10.1089/ast.2020.2342>.

Nardi, Luca. (2024) *“¿Cuáles Son Las Técnicas Utilizadas Para Descubrir Exoplanetas?”* WIRED. <https://es.wired.com/articulos/cuales-son-las-tecnicas-utilizadas-para-descubrir-exoplanetas>.

NASA, Equipo de redacción de Ciencia. (2022) *“Webb de La NASA Obtiene Su Primera Imagen Directa de Un Mundo Distante.”* NASA Ciencia. <https://ciencia.nasa.gov/universo/webb-de-la-nasa-obtiene-su-primera-imagen-directa-de-un-mundo-distante/>.

Nicolau, Jorge. (2025) *“Ra-Tp-Final.”* GitHub. <https://github.com/georgsmeinung/ra-tp-final/>.

Pimentel, J., Amorim, J. & Rudzicz, F. Feature extraction for exoplanet detection. *Int J Data Sci Anal* (2024). <https://doi.org/10.1007/s41060-024-00552-7>

Shields, Aomawa L., Ballard, Sarah, & Johnson, John Asher. (2016) “The Habitability of Planets Orbiting M-Dwarf Stars.” <https://doi.org/10.1016/j.physrep.2016.10.003>.

Tukey, J. W. (1977). Exploratory Data Analysis, Volumen 2 (18ª ed., ilustrada, reimpresa). Addison-Wesley Publishing Company.

Venkata, G & Jahnavi, M & Ch, Venkata & Suneetha, Muvva. (2023). Exoplanet Detection Using Feature Engineering with Ensemble Learning. <http://dx.doi.org/10.1109/ICPCSN58827.2023.00025>

Wolszczan, Aleksander & Frail, D. A. (1992) “A Planetary System Around the Millisecond Pulsar PSR1257 + 12.” <https://doi.org/10.1038/355145a0>.

Wu, Dong-Hong. (2023). *The possibility of detecting our solar system through astrometry*. arXiv.org. <https://arxiv.org/abs/2309.11729>