Analisis Multivariado de Exoplanetas: Identificación de Candidatos Habitables Utilizando PCA

Jose A. Govea-Garcia[1], Diego Vertiz-Padilla[1], Jose F. Gutierrez-Rubio[1], Gustavo A. Aguilar Torreblanca[1] and Emiliano Guzman-Ibarrola[1]

[1] *Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Guadalajara, Jalisco*

Abstract—

Keywords—Analisis de Componentes Principales, Exoplanetas, Analisis Multivariado, Reduccion de Dimensionalidad, Habitabilidad, Regresion lineal

I. Introducción

Un exoplaneta es un cuerpo celeste que se encuentra fuera del Sistema Solar y que orbita una estrella diferente al Sol. Dentro de esta categoría, los exoplanetas habitables son aquellos que se encuentran en la denominada "zona habitable" de un sistema planetario. Hasta la fecha, se han detectado más de 4,000 exoplanetas utilizando diversos métodos de observación, como las imágenes directas, que permiten bloquear el resplandor de las estrellas para detectar cuerpos celestes cercanos. Este avance ha planteado preguntas fundamentales, como la posibilidad de vida más allá de nuestro Sistema Solar o la viabilidad de encontrar un nuevo hogar para la humanidad en caso de una crisis planetaria. En este trabajo, se propone un análisis comparativo entre las características físicas de la Tierra y una base de datos de exoplanetas, con el objetivo de identificar las mejores opciones en habitabilidad. Para ello, se utilizara el Análisis de Componentes Principales (PCA), herramienta estadística que permite reducir la dimensionalidad de los datos, para facilitar su analisis.

II. Metodologia

Para poder realizar el análisis de los exoplanetas, seleccionamos un base de datos de parte de Nasa Exoplanet Archive que contiene un total del 36,490 observaciones (planetas) y 91 variables que describen a los exoplanetas, como lo son la masa, la temperatura, el radio del planeta, entre otros aspectos. Debido a que muchos de estas observaciones y variables no contaban con la información necesaria para poder realizar un análisis completo y apropiado, procedimos a realizar la limpieza de la base de datos. Utilizando el lenguaje de programacion Python, descartamos las observaciones y variables que no contaran con los suficientes datos y fueran intracendentes para realizar el analisis. Una vez realizada la limpieza de los datos, esta contaba con 1,182 observaciones y 17 variables con las cuales se realizaría el analisis.

Para realizar el analisis de la base de datos seguimos una serie de pasos importantes. De la misma forma que en la limpieza de la base de datos, se utilizó el lenguaje de programación Python y se trabajo en el entorno de Jupyter Notebooks. A la base de datos que limpiamos le quitamos las variables que no nos permiten realizar el analisis, como lo son el número del planeta y el nombre del planeta. Una vez realizado esto contamos con 15 variables las cuales describen las caracteristicas de los exoplanetas.

a. Matriz de Correlación

El primer analisis que se realizo fue con una matriz de correlación. En las matrices de correlación se observa la relaciones entre las distintas variables, ya sea que esten positivamente relacionadas, negativamente relacionadas o sean independientes entre si, con valores de 1 a -1 respectivamente. En caso de estar positivamente relacionadas, cuando la variable A aumenta, la variable B lo hace de la misma manera. Este es el caso de las variables st_teff (Temperatura Efectiva Estelar) y st_mass (Masa solar), que tienen un valor de 0.93. En caso de estar negativamente relacionadas, cuando la variable A aumente, la variable B va a disminuir. Este es el caso de las variables st_rad (Radio Estelar) y st_logg (gravedad de la superficie estelar). Por ultimo, encaso de ser independientes entre si, el valor será 0 o cercano a 0. Realizando este analisis se observa que varias variables esta relacionadas, ya sea de manera positiva o negativa. Esto es fundamental, ya que con esta información es posible y conveniente realizar un analisis de componentes principales, reduciendo la dimensiónalidad de los datos pero manteniendo la mayor cantidad de información posible.

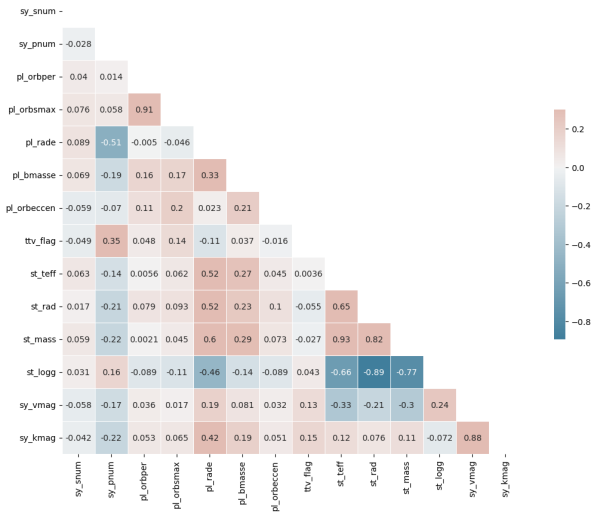

Fig. 1: Matriz de correlación que muestra la intensidad y dirección de las relaciones entre variables de exoplanetas.

b. PCA

Para realizar el analisis de componentes principales primero es fundamental estandarizár los datos a analizar. Ya que de no hacerlo, variables con escalas mas grandes tomarian mayor peso en los componentes principales. Con la base de datos estandarizada, utilizamos la biblioteca scikit-learn para implementar el PCA mediante un pipeline. En el pipeline se incluyo el ajuste de los datos con el modelo PCA. Con esto obtuvimos los componentes principales, los eigenvalores que muestran la cantidad de varianza en cada componente principal y los eigenvectores, que nos proporcionan la dirección de los componentes principales. Algunos componentes principales contienen muchas mas variabilidad que otros, es decir, mas informacion que otros.

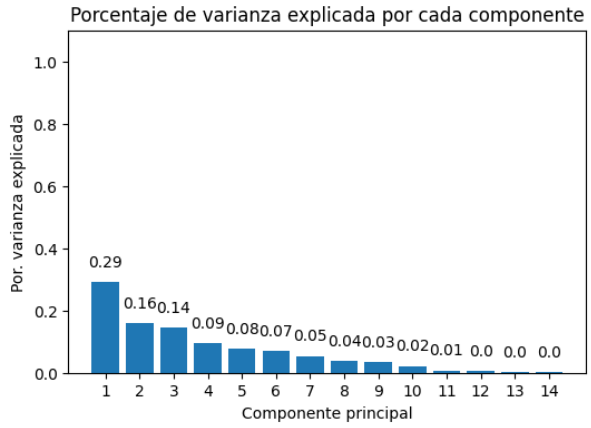

Fig. 2: Distribución del porcentaje de varianza explicada por cada componente principal

El objetivo de realizar un analisis de componentes principales es poder reducir la dimensiónalidad de los datos y representar estos en menos componentes, guardando la mayor cantidad de infromacion posible. Para poder determinar esto, es necesario visualiar la variablidad acomulada de los componentes y decidir la cantidad de componentes que mayor información almacenen. Tras esto se decidio que el numero adecuado de componentes principales es 9, ya que estos almacenan el 96% de los datos.

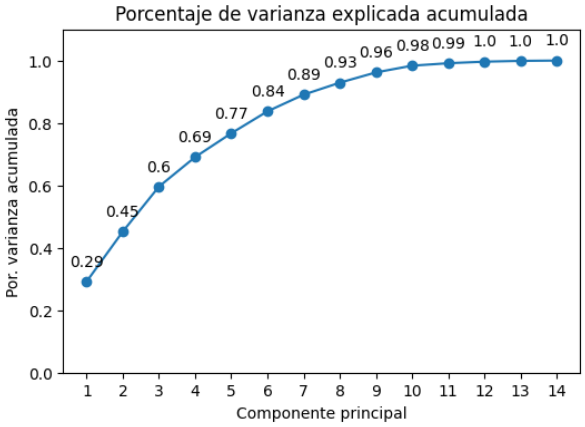

Fig. 3: Porcentaje de varianza acumulada por los componentes principales

Para tener certeza que el analisis de componentes principales se realizo de manera exitosa, realizamos una matriz de correlación entre los componentes principales. Con esta matriz de correlación observamos una relación perfecta en la diagonal, lo cual es cierto, ya que cada componente tiene una correlación de 1 consigo mismo. Mientras que la correlacion fuera de la diagonal observamos numeros muy puequeños, casi cercanos a cero. Esto nos indica que el analisis de componente principales se realizo de manera correcta. Ya que los componentes principales deben de ser ortogonales e independientes entre si.

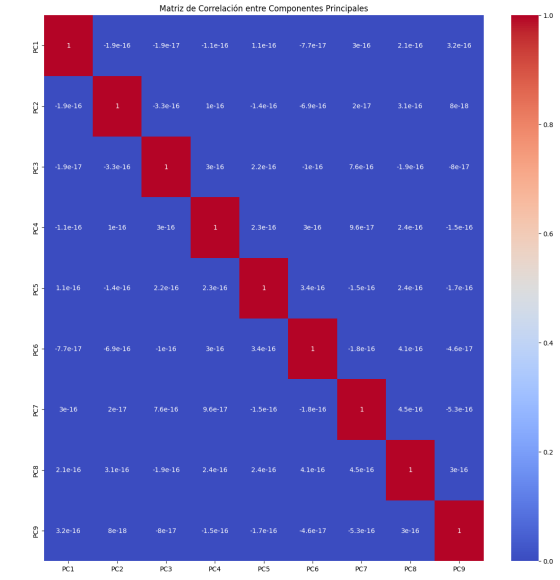

Fig. 4: Matriz de correlación entre componentes principales. Esto valida la independencia de los componentes generados por el PCA.

c. Vectores de Carga

Los vectores de carga son los coeficientes que representan en una combinación lineal a los componentes principales. Con esto podemos interpretar la información que contiene cada componente principal en el nuevo espacio respecto a las variables originales. Siendo los valores mas altos, positivos o negativos, los que tienen mayor influencia en el componente.

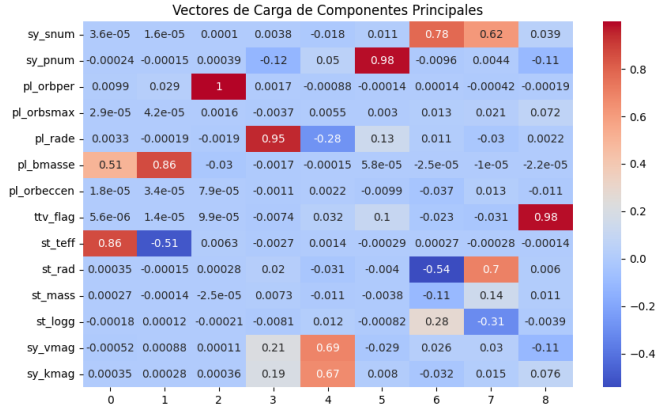

Fig. 5: Matriz de vectores de carga para los componentes principales. Permite interpretar la relación entre las variables originales y los componentes principales.

d. Análisis de Habitabilidad a partir de PCA

Posteriormente, con los resultados obtenidos por el Análisis de Componentes Principales, se creó una clase llamada ExoplanetHabitabilityAnalyzer, que cuenta con 4 métodos:

La primera función define la carga del componente. Esta define un peso a cada uno de los 9 componentes que describen la información. Esto basado en los vectores de carga que se mencionaron anteriormente. Es importante notar que para calcular el peso de cada uno de los componentes se tomó en cuenta un mapeo de las variables principales asociadas a cada componente con sus respectivas contribuciones y relevancias. Este cálculo se realizó mediante el producto de la varianza explicada por el componente, el valor absoluto de la contribución de cada variable líder dentro del componente y un puntaje de relevancia previamente asignado a cada variable. Este puntaje de relevancia se basa en un criterio establecido por Butturini, A. y colaboradores en su obra [1], donde se describen en detalle los factores que influyen en la determinación de la habitabilidad de cuerpos celestes.

- Componente Principal 1 (PC1): Con un peso de 0.4291 está dominado por las variables st_teff (Temperatura de la Estrella) y pl_bmasse (Masa Planetaria en términos de masas terrestres). Este componente captura las características físicas fundamentales, como la relación entre la masa planetaria y la temperatura estelar, factores esenciales para la habitabilidad.

- Componente Principal 2 (PC2): Se le asignó un peso de 0.2205 dado que este componente está liderado por pl_bmasse (positivamente) y st_teff (negativamente). Representa el contraste entre la masa planetaria y el ambiente estelar, resaltando cómo estos factores interactúan para influir en la habitabilidad potencial.

- Componente Principal 3 (PC3): Tiene asignado 0.0741 como peso, porque está dominado casi exclusivamente por la variable pl_orbper (Período Orbital). Representa el ciclo temporal del planeta, asociado a su posición en la zona habitable de la estrella.

- Componente Principal 4 (PC4): Este componente está dominado por pl_rade (Radio Planetario) y, en menor medida, sy_vmag (Magnitud Visible) y tiene un peso de 0.1328. Representa el tamaño del planeta y su visibilidad, cruciales para evaluar planetas detectables y habitables.

- Componente Principal 5 (PC5): COn peso de 0.0541, está liderado por sy_vmag y y_kmag (Magnitudes en bandas visibles e infrarrojas). Describe las propiedades fotométricas, como la luminosidad del sistema.

- Componente Principal 6 (PC6): Este componente está dominado por sy_pnum (Número de Planetas en el Sistema) y se le asignó 0.00889 como peso. Refleja la multiplicidad del sistema planetario, que influye en su dinámica orbital.

- Componente Principal 7 (PC7): Está liderado por sy_snum (Número de Estrellas en el Sistema) y st_rad (Radio de la Estrella) y tiene peso de 0.0654. Representa la arquitectura estelar, relevante para la estabilidad orbital.

- Componente Principal 8 (PC8): Este componente combina varios factores con pesos moderados, por lo que toma un peso de 0.0025. Representa las características mixtas del sistema planetario.

- Componente Principal 9 (PC9): Tiene un peso de 0.0126 y está dominado por ttv_flag (Variaciones Temporales en el Tránsito Planetatio). Representa las irregularidades orbitales, que pueden aportar información adicional sobre la dinámica del sistema.

Para evaluar la habitabilidad de los exoplanetas, se definieron rangos óptimos para cada uno de los 9 Componentes Principales (PC). Estos rangos reflejan valores que, según el análisis y la literatura científica, representan condiciones favorables para la habitabilidad.

La clase utiliza un diccionario llamado optimal_ranges, que contiene los valores mínimos y máximos para cada componente principal. La lógica detrás de estos rangos es la siguiente:

- PC1 (Características Físicas Fundamentales): Este rango evalúa un balance entre la masa planetaria y la temperatura estelar. Un rango moderado (min = -1.5, max = 1.5) asegura que los planetas no sean extremos en estas características.

- PC2 (Masa vs Temperatura): Un rango óptimo (min = -1.0, max = 1.0) favorece planetas con una combinación razonable de masa y temperatura.

- PC3 (Período Orbital): Los períodos orbitales moderados (min = -2.0, max = 2.0) indican órbitas estables dentro de la zona habitable.

- PC4 (Tamaño del Planeta): Un rango cercano a 0 (min = -1.0, max = 1.0) favorece planetas de tamaño similar a la Tierra.

- PC5 (Luminosidad del Sistema): Rango amplio (min = -2.0, max = 2.0) para incluir una variedad de sistemas con luminosidades moderadas.

- PC6 (Multiplicidad del Sistema): Un rango limitado (min = -1.0, max = 1.0) refleja sistemas con una estructura planetaria no demasiado compleja.
- PC7 (Estabilidad Estelar): Estrellas estables y bien equilibradas tienen valores en el rango (min = -1.5, max = 1.5).
- PC8 (Características Mixtas): Un rango moderado (min = -1.0, max = 1.0) captura combinaciones de características sin favorecer extremos.
- PC9 (Órbita Estable): Se busca estabilidad orbital en el rango (min = -1.0, max = 1.0).

Estos rangos óptimos son utilizados por la función calculate_pc_score para calcular los puntajes individuales de cada componente principal, normalizando los valores de cada planeta y determinando qué tan cerca se encuentran de las condiciones ideales.

La segunda función calcula el puntaje de cada componente principal. El método es calculate_pc_score, que calcula un puntaje para cada componente principal basado en los valores proyectados en el PCA. Este puntaje evalúa qué tan bien se encuentra cada componente dentro de un rango óptimo previamente definido. El cálculo incluye los siguientes pasos:

1. Se toma el valor del componente principal para un planeta específico.
2. Se normaliza este valor al rango óptimo definido para ese componente principal, utilizando la fórmula:

$$\texttt{normalized\_value} = \frac{\texttt{pc\_value} - \texttt{min}}{\texttt{max} - \texttt{min}}$$

3. Se calcula el puntaje, asegurándose de que esté entre 0 y 1. Este valor representa qué tan "bueno" es el valor proyectado del componente principal con respecto a los valores ideales:

$$\texttt{score} = 1 - \min(1, |\texttt{normalized\_value} - 0.5| \times 2)$$

La tercera función calcula el puntaje de habitabilidad. El método calculate_habitability_score evalúa la habitabilidad general de un exoplaneta al combinar los puntajes individuales de los Componentes Principales (PC) en una única métrica ponderada. Este cálculo se realiza de la siguiente manera:

1. Para cada Componente Principal (PC1 a PC9), se utiliza la función calculate_pc_score para calcular un puntaje individual. Este puntaje mide qué tan cerca está el valor del PC de un rango considerado óptimo para la habitabilidad.
2. Cada puntaje individual es multiplicado por el peso asignado al correspondiente PC en el diccionario pca_weights. Estos pesos reflejan la importancia relativa de cada PC en la evaluación de la habitabilidad.
3. Los puntajes ponderados se suman para obtener el puntaje total de habitabilidad del exoplaneta.

El resultado final es un puntaje entre 0 y 1, donde 1 representa condiciones ideales de habitabilidad según los rangos y pesos definidos. Este puntaje permite clasificar a los exoplanetas en términos de su potencial para ser habitables, considerando múltiples factores físicos y orbitales.

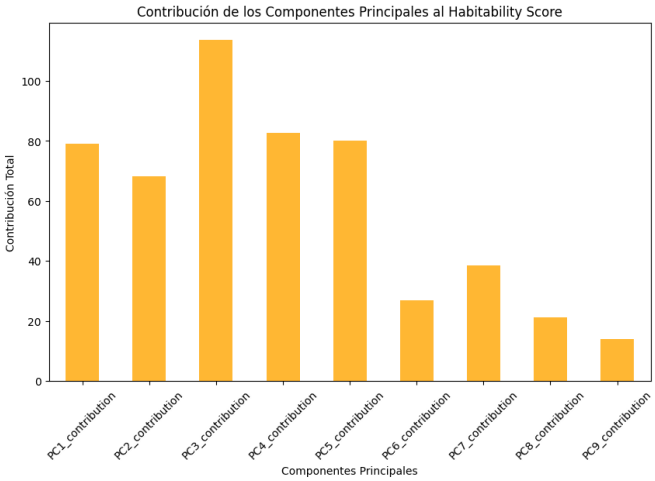

Fig. 6: Contribución de los componentes principales al puntaje de habitabilidad.

La cuarta funciona realiza el analisis completo de habitabilidad. El método analyze realiza un análisis exhaustivo de los exoplanetas basado en los datos originales y sus proyecciones reducidas por PCA. Su funcionamiento incluye los siguientes pasos:

1. Validación de datos: Se verifica que el número de filas en los datos originales (X) y las proyecciones PCA (pca_transformed) sea el mismo. Además, se asegura que las proyecciones incluyan al menos 9 Componentes Principales.
2. Cálculo del puntaje de habitabilidad: El método invoca calculate_habitability_score para calcular el puntaje total de habitabilidad de cada exoplaneta en el conjunto de datos.
3. Clasificación de habitabilidad: Los puntajes de habitabilidad se agrupan en cinco categorías (Very Low, Low, Moderate, High, Very High) utilizando el método qcut de pandas, que divide los datos en quintiles.
4. Contribuciones de cada PC: Para cada Componente Principal, se calcula su contribución individual al puntaje total de habitabilidad. Estas contribuciones se añaden como columnas al DataFrame de resultados.

El método devuelve un DataFrame con las siguientes columnas:

- habitability_score: El puntaje total de habitabilidad.
- habitability_class: La categoría de habitabilidad basada en el quintil.
- PCn_contribution: La contribución específica de cada Componente Principal al puntaje total, donde n varía de 1 a 9.

Este análisis permite identificar no solo qué planetas son más habitables, sino también los factores clave que contribuyen a su habitabilidad.

Por ultimo, la ultima función selecciona los mejores candidatos. El método `get_top_candidates` identifica a los exoplanetas con mayor potencial de habitabilidad al seleccionar los mejores `n` candidatos basados en sus puntajes de habitabilidad. Este proceso se lleva a cabo de la siguiente manera:

1. Se utiliza el método nlargest de pandas para ordenar los datos según la columna habitability_score y seleccionar los n planetas con los puntajes más altos.

2. Se combinan las columnas relevantes del conjunto de datos original con las columnas clave del conjunto de resultados, como el puntaje y la clasificación de habitabilidad.

3. Se seleccionan las siguientes características principales para presentar en el reporte:

    - pl_name: Nombre del exoplaneta.
    - habitability_score: Puntaje total de habitabilidad.
    - habitability_class: Clasificación de habitabilidad (e.g., ery High, High, Moderate).
    - pl_rade: Radio planetario (en radios terrestres).
    - pl_bmasse: Masa planetaria (en masas terrestres).
    - pl_orbper: Período orbital (en días).
    - st_teff: Temperatura efectiva de la estrella (en Kelvin).

El resultado final es un DataFrame que presenta un resumen detallado de los mejores candidatos para habitabilidad. Esto permite priorizar los exoplanetas más prometedores para futuros estudios y observaciones.

e. Regresión Lineal

Con el propósito de complementar el análisis PCA previamente realizado, es empleado el método de regresión lineal múltiple, una técnica de modelado que permite analizar datos y hacer predicciones. El modelo de regresión múltiple relaciona las variables independientes $x_1, x_2, ..., x_n$ y una variable dependiente y observadas en una muestra de datos dada. También protege de realizar conclusiones imprecisas basadas en correlaciones parciales. Para la realización del modelo es necesario establecer de manera clara las variables dependiente e independientes que componen el modelo para después determinar los coeficientes de cada x (cita 1).

La regresión múltiple se utilizó en los datos originales y en el vector de habitabilidad obtenido anteriormente, con el objetivo de crear un modelo que pueda predecir la habitabilidad de un planeta. Es fundamental que las variables independientes sean linealmente independientes entre si, ya que de otra forma aportan información similar. Estas variables redundantes, al proporcionar información estrechamente relacionada al sistema, provocan inestabilidad y se vuelven susceptibles a pequeñas fluctuaciones en los datos (cita ?). Al observar la figura 1, se puede visualizar que la correlación entre las variables es en su mayoría entre baja e intermedia, por lo que éstas son, potencialmente, linealmente independientes.

Al no haber observado claros indicios de dependencia lineal en la figura 1, se computó un modelo de regresión múltiple utilizando el conjunto de datos. Tras generar el modelo se observó un coeficiente de $R^2 = 0.416$, que sugiere que el modelo describe solamente el 41.6% de los datos. El modelo generado cuenta con un intercepto en el eje y diferente al origen, por lo que se se generó otro modelo sin intercepto, buscando aumentar el coeficiente de R^2. El segundo modelo describe un 89.2% de los datos, por lo que se optó utilizar éste.

III. Resultados

En términos de habitabilidad, después del cálculo según los Componentes Principales, la función `get top candidates` arrojó que los 10 planetas con mejor Puntaje de Habitabilidad fueron los siguientes:

Planeta	Puntaje de Habitabilidad	Clase de Habitabilidad
HATS-38 b	0.9342	Very High
TOI-2196 b	0.8364	Very High
WASP-41 b	0.8346	Very High
WASP-105 b	0.8291	Very High
WASP-130 b	0.8279	Very High
TOI-905 b	0.8271	Very High
TOI-1420 b	0.8113	Very High
WASP-47 b	0.7970	Very High
WASP-60 b	0.7964	Very High
TOI-2000 c	0.7917	Very High

Fig. 7: Top 10 planetas con mayor Puntaje de Habitabilidad

- HATS-38 b (Puntaje 0.93), TOI-2196 b (Puntaje 0.84): Planetas descubiertos mediante misiones de búsqueda de exoplanetas que amplían nuestra comprensión de sistemas planetarios diversos.

  Fuente de investigación:

  – Espinoza, N., et al. (2022). “Comprehensive Exoplanet Surveys: Methods and Discoveries”. Annual Review of Astronomy and Astrophysics, 60, 187-224.

- WASP-41 b (Puntaje 0.83), WASP-105 b (Puntaje 0.83), WASP-130 b (Puntaje 0.83), WASP-47 b (Puntaje 0.80), WASP-60 b (Puntaje 0.80): Exoplanetas identificados en el proyecto WASP, representativos de la categoría de “Júpiter calientes”.

  Fuente de investigación:

  – Hellier, C. (2018). “The WASP Planetary Transit Survey”. Exoplanet Science, Cambridge University Press.

- TOI-905 b (Puntaje 0.83), TOI-1420 b (Puntaje 0.81), TOI-2000 c (Puntaje 0.79): Planetas detectados por la misión TESS, proporcionando nuevas perspectivas sobre la diversidad planetaria.

  Fuente de investigación:

  – Ricker, G. R., et al. (2021). “TESS Mission: First Four Years of Exoplanet Discoveries”. The Astronomical Journal, 162(3), 95.

A continuación, se presenta el gráfico que compara los exoplanetas en términos de los Componentes Principales 1 y 2, que son los que explican la mayor parte de la variabilidad en los datos. En este espacio, se observa una mayor dispersión y menos agrupamiento, lo que resalta las diferencias entre los exoplanetas con los mejores puntajes de habitabilidad.

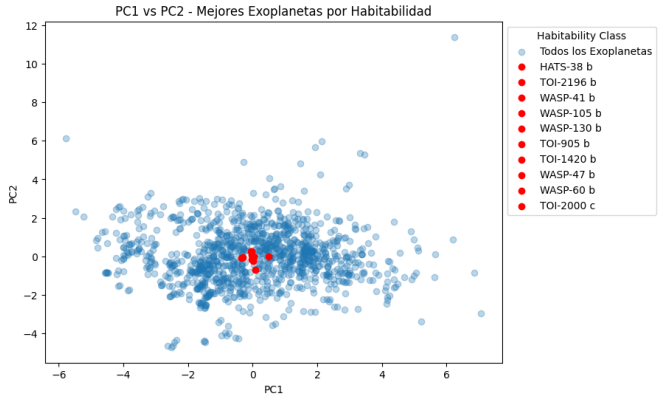

Fig. 8: Distribución de exoplanetas en el espacio de Componentes Principales: Análisis de características de habitabilidad

El análisis de componentes principales revela una distribución informativa de exoplanetas con potencial de habitabilidad. La representación gráfica muestra una dispersión significativa en los dos primeros componentes, lo que evidencia la complejidad y variabilidad de las características planetarias estudiadas.

La posición central de estos exoplanetas en el espacio de componentes sugiere condiciones intermedias pero prometedoras. Esta ubicación indica que no presentan valores extremos, sino características que podrían configurar entornos potencialmente estables para procesos compatibles con la vida.

REFERENCES

[1] A. Butturini, D. García-Castellanos, C. Jordi, I. Ribas, and J. Urmeneta, *(In)habitabilidad planetaria*. Marcombo, 2020.