CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

INFOTEC

BIBLIOTECA INFOTEC
VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

Ciudad de México, a 12 de noviembre de 2024

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Los exoplanetas son como caramelos cósmicos: Cómo clasificar una amplia gama de sabores con aprendizaje automático"

Desarrollado por la alumna: **Monserrat Campos Sánchez,** bajo la asesoría del **Dr. Miguel Ángel Porta García** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que la aluna no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Mtro. Carlos Josué Lavandeira Portillo
Director Adjunto de Innovación y Conocimiento

CJLP/dqor

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento.
Monserrat Campos Sánchez.- Alumna de la Maestría en Ciencia de Datos e Información.- Para su conocimiento.

MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

“Los exoplanetas son como caramelos cósmicos: Cómo clasificar una amplia gama de sabores con aprendizaje automático”

IMPLEMENTACIÓN DE PROYECTO LABORAL

Que para obtener el grado de MAESTRA EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Monserrat Campos Sánchez

Asesor:

Dr. Miguel Angel Porta García

Ciudad de México, Mayo, 2024.

CONAHCYT

INFOTEC POSGRADOS

Agradecimientos

Empiezo este trabajo agradeciendo a todos mis maestros de cualquier etapa, en especial aquellos que aman lo que hacen y lo saben compartir. Gracias por hacer mi mundo más grande.

Quiero dedicar este trabajo a mi familia. Las palabras no expresan la admiración que siento por cada uno de ustedes, veo cómo luchan y se esfuerzan por ser mejores personas cada día. Mi familia representa mis raíces que brindan soporte y nutrientes, no todo el tiempo son visibles, pero existen estos pequeños momentos en los que puedo nombrarles y mostrarles mi admiración, cariño y respeto. Gracias por anclarse tan fuerte y permitirme crecer.

Finalmente, quisiera que este trabajo sea un recordatorio para mí de las grandes cosas que puedo lograr, pero no solo eso, recordar las dificultades enfrentadas y que sean un motivo de orgullo, pidiendo a mi niña interior que siga soñando para conseguir y luchar por logros y metas más grandes.

Índice general

Índice de figuras

Índice de tablas

Abreviaturas y acrónimos

NASA	Administración Nacional de Aeronáutica y el Espacio (del inglés "National Aeronautics and Space Administration").
SMOTE	Técnica de Sobremuestreo de Minorías Sintéticas (del inglés "Synthetic Minority Over-sampling Technique").
AU	Unidad Astronómica (del inglés, “Astronomical Unit”)
RA	Ascensión Recta (del inglés, “Right Ascension”)
PCA	Análisis de Componentes Principales (del inglés, “Principal Component Analysis”)
CCD	Dispositivo de Carga Acoplada (del inglés, “Charge Coupled Device”)
KDD	Descubrimiento de Conocimiento en Bases de Datos (del inglés, Knowledge Discovery in Databases)
SVM	Máquinas de Soporte Vectorial (del inglés, Support Vector Machine)

Resumen

Con el descubrimiento de miles de exoplanetas en las últimas décadas, surge la necesidad de desarrollar métodos efectivos para clasificarlos y caracterizarlos. Este estudio se centra en la clasificación de exoplanetas al implementar Redes Neuronales usando datos obtenidos de la misión Kepler.

Se establece una base de datos preprocesada que incluye variables clave del sistema planetario, lo que permite mejorar la precisión y eficiencia en la clasificación. Se describe el proceso de recopilación de datos y las etapas de preprocesamiento necesarias para preparar los datos para el análisis. Esto incluye la limpieza de datos, la manipulación de datos faltantes y la transformación de variables categóricas en numéricas.

La importancia de esta investigación radica en comprender la diversidad de exoplanetas, identificar su habitabilidad y apoyar futuras misiones espaciales, destacando la sinergia interdisciplinaria en este campo.

Además, se explora el procedimiento de ajuste de parámetros en la Red Neuronal, donde estos influyen significativamente en el rendimiento y la precisión de los resultados. En este proceso, cada parámetro se evalúa en una ronda, donde se configuran los valores por defecto de cada parámetro y solo para un parámetro se prueban diferentes valores. Para evaluar los parámetros y evitar el sobreentrenamiento se hace uso de la validación cruzada, donde la mejor precisión es seleccionada y pasa a la siguiente ronda.

En cuanto a las métricas de Sensibilidad, Tasa Negativa Verdadera, Valor Predictivo Positivo y Valor Predictivo Negativo, el Tipo Terrestre mostró consistentemente las probabilidades más altas para todas estas métricas, seguido por el Tipo Gigante Gaseoso, luego Súper Tierra y finalmente Tipo Neptuno.

Capítulo 1. Generalidades

En este primer capítulo se expone el planteamiento del problema donde se busca definir la problemática, así como su protocolo de investigación donde se expondrán las preguntas, objetivos e hipótesis que persigue esta investigación. En la sección de justificación se busca informar al lector sobre la importancia del estudio y análisis de la problemática, y finalmente en la sección de los límites y alcances se busca definir la problemática de esta investigación, centrando así la investigación.

1.1 Planteamiento del problema.

Los planetas, cuerpos celestes de diversos tamaños con forma redondeada, formados a partir de la misma materia gaseosa que dio origen a las estrellas o que provino de ellas, con una constitución infinitamente más fría y comprimida, adquirieron así propiedades físicas y químicas distintas (Etecé, 2021). Cuando estos orbitan estrellas diferentes a la nuestra, conocido como Sol, se les denomina exoplanetas o planetas extrasolares (Ruíz, 2017). De manera general, los planetas pueden ser clasificados en cuatro tipos: Gigante Gaseoso, Súper Tierra, Tipo Neptuno y Terrestre (González, 2021). La mayor complejidad al estudiar los exoplanetas se encuentra en las grandes distancias, donde ya ni siquiera es posible mirarlos directamente con telescopios, y algunos otros se encuentran ocultos por el resplandor de estrellas brillantes; como resultado, el ser humano ha enviado misiones como KEPLER para su estudio, generado así una gran cantidad de datos. Comprender los tipos de planetas constituye un desafío fundamental en la astronomía contemporánea debido a la importante influencia que tiene en la comprensión sobre la diversidad del universo y la identificación de la habitabilidad. La clasificación de los exoplanetas es una tarea compleja debido a la gran cantidad y naturaleza de los datos, y además la necesidad de algoritmos capaces de discernir y catalogar estos diferentes cuerpos celestes.

A medida que se ha incrementado el estudio de la astronomía y el uso de misiones como KEPLER, el tamaño y complejidad de bases de datos astronómicos han ido en aumento, brindando información sobre la diversidad de exoplanetas, así como su clasificación. Sin embargo, esta gran cantidad de datos hace que disminuya la capacidad de los astrónomos para buscar datos manualmente y plantea desafíos como la complejidad y necesidad de técnicas de análisis efectivas (Giles, 2018).

Ball (2010) afirma que la minería de datos, si se usa correctamente puede tener un enfoque poderoso, ya que tiene el potencial de explorar cantidades de datos cada vez mayores utilizando algoritmos de ciencia de datos que contribuyen a la mejora de la astronomía; prometiendo así un gran avance científico. Aprovechando el poder del aprendizaje automático, se pueden entrenar algoritmos para descifrar patrones en los datos y desarrollar modelos capaces de clasificar exoplanetas en función de los atributos observados (Prithivraj, 2023). Las variables del sistema planetario de la misión KEPLER ofrecen una oportunidad única de explorar las relaciones entre estas variables y los tipos de exoplanetas que caracterizan.

El proceso de implementación y ajuste de algoritmos de ciencia de datos presenta la necesidad de un análisis metódico. Por lo que se deben seleccionar algoritmos apropiados, así como su ajuste de hiperparámetros e integrar técnicas de ingeniería de características para determinar la precisión y confiabilidad de los resultados de la clasificación. La investigación de estos factores contribuye no solo al campo de la investigación de exoplanetas, sino también a la aplicación más amplia de metodologías de ciencia de datos.

En vista de estas consideraciones, esta investigación tiene como objetivo explorar el potencial de los algoritmos de ciencia de datos para clasificar tipos de exoplanetas en función de variables del sistema planetario obtenidas de la misión

KEPLER. Al evaluar diferentes algoritmos, ajustar sus parámetros y examinar el impacto de la selección de variables y la ingeniería de características, este estudio busca mejorar nuestra capacidad para clasificar exoplanetas con precisión, proporcionando una comprensión más profunda de su diversidad y contribuyendo al avance de la investigación de exoplanetas.

1.2 Protocolo de investigación.

En la Tabla 1 se exponen las preguntas, objetivos e hipótesis de esta investigación. Las preguntas definen el enfoque de la investigación, por lo que mostrará las principales áreas de trabajo en esta investigación, siendo la construcción de la base de datos, ajuste e implementación de los algoritmos y evaluación de los algoritmos. En el caso de los objetivos, se establecen las metas de la investigación y finalmente las hipótesis que son las predicciones que se tienen sobre los resultados de esta investigación.

	Pregunta	**Objetivo**	**Hipótesis**
General	¿Cómo se ajustarían y usarían los algoritmos de aprendizaje computacional para clasificar los tipos de exoplanetas usando las variables del sistema planetario obtenidas de la misión KEPLER?	Clasificar los tipos de exoplanetas mediante la aplicación de algoritmos de aprendizaje computacional utilizando variables del sistema planetario obtenidas de la misión KEPLER.	Los algoritmos de aprendizaje computacional, con el ajuste adecuado, se pueden usar para clasificar de manera efectiva los tipos de exoplanetas utilizando las variables del sistema planetario obtenidas de la misión KEPLER. Al aplicar técnicas estadísticas y de aprendizaje automático a este rico conjunto de datos, podemos identificar y analizar las variables más relevantes para la clasificación de exoplanetas y desarrollar modelos que predigan con precisión el tipo de exoplaneta en función de estas variables.

Específicos	1	¿Cuáles son las variables clave del sistema planetario que deben identificarse y seleccionarse para construir una base de datos de exoplanetas preprocesada?	Construir una base de datos de exoplanetas preprocesada mediante la identificación y selección de variables clave del sistema planetario.	Al identificar y seleccionar variables clave del sistema planetario, se puede construir una base de datos de exoplanetas preprocesada, lo que puede mejorar la precisión y la eficiencia de la clasificación de exoplanetas.
	2	¿Cómo se pueden ajustar y aplicar los algoritmos de aprendizaje computacional para clasificar los tipos de exoplanetas utilizando variables del sistema planetario?	Implementar y ajustar algoritmos de aprendizaje computacional utilizando la base de datos de exoplanetas creada, para clasificar los tipos de exoplanetas en función de las variables del sistema planetario.	Al implementar y ajustar algoritmos de aprendizaje computacional utilizando la base de datos de exoplanetas creada, es posible clasificar con precisión los tipos de exoplanetas en función de las variables del sistema planetario.
	3	¿Cómo podemos evaluar el rendimiento de los algoritmos de aprendizaje computacional para la clasificación de exoplanetas usando métricas relevantes y qué podemos aprender del análisis y la comparación de los resultados?	Evaluar el desempeño de los algoritmos de aprendizaje computacional para la clasificación de exoplanetas analizando y comparando los resultados utilizando métricas relevantes.	El rendimiento de los algoritmos de aprendizaje computacional para la clasificación de exoplanetas se puede evaluar utilizando métricas relevantes, y el análisis y la comparación de los resultados pueden proporcionar información sobre la efectividad de los diferentes algoritmos y la importancia de las diferentes variables del sistema planetario.

Tabla 1. Protocolo de Investigación.

Fuente: Elaboración propia.

1.3 Justificación.

Esta investigación busca, como menciona Prithivraj (2023) aprovechar el poder del aprendizaje automático para entrenar algoritmos capaces de descifrar patrones en los datos y desarrollar modelos capaces de clasificar exoplanetas en función de

los atributos observados. Y se pretende que mientras se realiza esta investigación se pueda ayudar a comprender la diversidad de exoplanetas, identificar la habitabilidad, favorecer a futuras misiones espaciales, sinergia interdisciplinaria y búsqueda de conocimiento.

Como primer motivo, se tiene la comprensión sobre la diversidad de exoplanetas, es decir, se puede apoyar a fomentar el conocimiento acerca de cómo son los procesos de formación y evolución en un planeta, dando lugar a información sobre los mecanismos que se dan en diferentes tipos de planetas. Además, los planetas existen con una amplia gama de tamaños, composiciones y entornos; por lo que se puede explorar las distintas condiciones de los planetas, generando conocimiento de los ambientes y su potencial de vida en estos.

El segundo motivo se relaciona con la identificación de la habitabilidad. Siempre ha existido una pregunta que persigue al ser humano: ¿existe vida fuera de nuestro planeta? El poder clasificar exoplanetas ayuda a generar objetivos específicos para la búsqueda de vida. Además, se pueden encontrar cuáles son los límites para que pueda existir vida, es decir, se pueden estudiar cuáles son las condiciones bajo las cuales se puede dar la vida. Finalmente, puede aportar información sobre candidatos habitables y entendimiento de la singularidad de la Tierra para permitir la vida.

El tercer motivo es favorecer a futuras misiones espaciales. Se debe recordar que los recursos son limitados y las misiones son costosas, por lo que tener una lista de candidatos potenciales ayuda a la selección de las misiones optimizando recursos y asegurar que cada misión sea efectiva. A su vez, permite la preparación de las misiones al adaptarlas de acuerdo a los entornos físicos-químicos de los diferentes tipos de exoplanetas; optimizando así las técnicas de recopilación de datos, permitiendo adaptar los sensores e instrumentos, generando así observaciones más precisas y detalladas.

El cuarto motivo es la sinergia interdisciplinaria, en donde se encuentra que la colaboración de diferentes campos, como lo son la astronomía, ciencia de datos, astrofísica y biología, pueden dar una comprensión más integral sobre los factores que influyen en los tipos de exoplanetas. A su vez, a medida que se ha incrementado el estudio de la astronomía y el uso de misiones como KEPLER, el tamaño y la complejidad de bases de datos astronómicos han ido en aumento, brindando información sobre la diversidad de exoplanetas, así como su clasificación. Sin embargo, esta gran cantidad de datos hace que disminuya la capacidad de los astrónomos para buscar datos manualmente y plantea desafíos como la complejidad y necesidad de técnicas de análisis efectivas (Giles, 2018).

Finalmente, el último motivo es la necesidad del ser humano por la búsqueda de conocimiento. El ser humano siempre ha mirado hacia arriba y se ha preguntado: ¿qué hay más allá? Esta investigación busca aportar un poco a esa pregunta, volviendo factible la exploración. Esto genera nuevas técnicas de observación, análisis de datos y modelos teóricos, conduciendo a nuevos descubrimientos y avances para la ciencia y tecnología.

En conclusión, la clasificación de los tipos de exoplanetas contribuye al progreso en varios frentes. Informa nuestra comprensión de la formación planetaria, la habitabilidad y el potencial de vida más allá de la Tierra. Este conocimiento puede tener implicaciones de gran alcance para campos como la astrobiología, las ciencias planetarias y la astronomía.

1.4 Límites y alcances.

1.4.1 Límites.

El primer límite al que se enfrenta la investigación es la calidad y cantidad de datos; como en esta investigación solo se usarán datos recolectados en la misión Kepler, se depende de su calidad y cantidad; así mismo se enfrentan con

imprecisiones, datos faltantes y limitados que pueden afectar al rendimiento de los algoritmos de ciencia de datos.

Una segunda limitación que enfrenta esta investigación es la complejidad de los tipos de exoplanetas; esto se debe a su naturaleza complicada y confusa. En los datos se pueden encontrar características ambiguas que sean difíciles de clasificar.

Finalmente, otro límite está determinado por la generalización de los algoritmos. Si se obtienen buenos resultados mediante algoritmos específicos utilizando estrategias de ajuste, se debe analizar hasta qué punto los resultados pueden ser generalizados a otro conjunto de datos; es decir, puede existir una relación directa entre la efectividad de los algoritmos y características específicas de los datos.

1.4.2 Alcances.

Dentro de la investigación será posible implementar y ajustar diferentes tipos de algoritmos de aprendizaje computacional para la clasificación del tipo de exoplaneta. Una vez implementados estos algoritmos serán evaluados mediante métricas relevantes para finalmente ser comparados basados en el rendimiento, esta información mostrará que algoritmos son los más adecuados y en qué condiciones; por lo tanto, un primer alcance será la diversidad de algoritmos y comparación del rendimiento.

El impacto en la selección de variables será otro alcance de la investigación, ya que se explorarán diferentes combinaciones de variables del sistema planetario, buscando si existe alguna combinación que mejore la calidad de los resultados a la hora de clasificar el tipo de exoplaneta.

Finalmente, el último alcance de esta investigación es el ajuste competitivo para los algoritmos de aprendizaje computacional; para lograr este alcance se

requerirán diferentes estrategias donde se exploran los efectos de la selección de hiperparámetros, ingeniería de características y técnicas de preprocesamiento.

Capítulo 2
Base de datos

Capítulo 2. Base de datos

Este segundo capítulo tiene como finalidad exponer la base de datos construida, la cual contendrá las variables del sistema planetario que alimentarán la clasificación del tipo de exoplaneta. El proceso para poder construir la base de datos implica encontrar los datos desde diferentes fuentes de información, para así recolectarlos y generar las variables. Una vez que se tienen los datos, es necesario realizar un preprocesamiento que ayude a limpiar los datos, manipular datos faltantes y transformar las variables categóricas en numéricas. Finalmente, se realiza un análisis exploratorio de datos que permite entender cómo están construidas las variables, así como su relación con respecto a otras. En cada sección del capítulo se describen la aplicación de estos pasos para la base de datos utilizada en la investigación.

2.1 Construcción de la base de datos

En 2009, la Administración Nacional de Aeronáutica y el Espacio, conocida como NASA (por sus siglas en inglés, National Aeronautics and Space Administration) en colaboración con otras instituciones, lanzaron el Telescopio Espacial Kepler, el cual tenía entre sus objetivos buscar planetas usando el método de tránsitos planetarios (Rojas, 2016). En esta misión se buscaron planetas de varios tamaños y órbitas, y estos a su vez orbitaban alrededor de estrellas de distinto tamaño y temperatura. Como resultado de los cuatro años de la misión, se descubrieron varios miles de planetas candidatos observando más de 150,000 estrellas en la constelación de Cygnus-Lyra (Furlan et al., 2017).

2.1.1 Fuentes de información

El Archivo de exoplanetas de la NASA es un servicio de datos y catálogo estelar y de exoplanetas astronómicos en línea que recopila y correlaciona datos astronómicos e información sobre exoplanetas y sus estrellas anfitrionas, y proporciona herramientas para trabajar con estos datos (Alexander, 2021). Se incluyen parámetros estelares, parámetros de exoplanetas y datos de descubrimiento/caracterización.

Por otro lado, la enciclopedia exoplanetaria de la NASA es continuamente actualizada combinando visualizaciones interactivas con datos detallados sobre todos los exoplanetas conocidos. En esta base de datos se encuentran solo los planetas confirmados a diferencia de la primera y solo posee información del nombre del exoplaneta, años luz de la tierra, masa del planeta, magnitud estelar, la fecha de descubrimiento y el tipo de planeta.

2.1.2 Creación de la base de datos

Utilizando la Enciclopedia Exoplanetaria (link de la página https://science.nasa.gov/exoplanets/exoplanet-catalog/) se enlistan todos los exoplanetas de la misión “Kepler/K2” tomando el nombre y tipo de planeta. A continuación, utilizando la segunda fuente de información, es decir, el Archivo de exoplanetas de la NASA se utiliza la tabla de “Planetary System” (dicha tabla puede ser accedida a través de esta dirección https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS), donde se muestra todas las soluciones para planetas y estrellas anfitrionas, independientemente de su relación; esto incluye sistemas atípicos, como planetas que flotan libremente y aquellos con múltiples estrellas. Esta tabla también contiene soluciones candidatas de Kepler, K2 y TESS

para sistemas planetarios confirmados, una identificación casi completa de compañeros estelares publicados y oblicuidades planetarias reales y proyectadas. Dentro de la tabla se recopilarán las variables del sistema planetario y se adjuntarán al nombre y tipo de planeta.

Para esta investigación se utilizarán datos recopilados por la misión KEPLER con alrededor de 3253 entradas. Con las siguientes características:

- Nombre del planeta
- Tipo de planeta
- Número de estrellas en el sistema planetario.
- Número de planetas en el sistema planetario.
- Tiempo que tarda el planeta en realizar una órbita completa alrededor de la estrella o sistema anfitrión
- El radio más largo de una órbita elíptica o, para los exoplanetas detectados mediante microlente gravitacional o imágenes directas, la separación proyectada en el plano del cielo.
- Longitud de un segmento de línea desde el centro del planeta hasta su superficie, medida en unidades de radio de la Tierra
- Longitud de un segmento de línea desde el centro del planeta hasta su superficie, medida en unidades de radio de Júpiter
- Masa mínima de un planeta medida por la velocidad radial, medida en unidades de masa de la Tierra
- Masa mínima de un planeta medida por la velocidad radial, medida en unidades de masa de Jupiter
- Cantidad por la cual la órbita del planeta se desvía de un círculo perfecto
- El flujo de insolación es otra forma de dar la temperatura de equilibrio. Se da en unidades relativas a las medidas de la Tierra desde el Sol.
- La temperatura de equilibrio del planeta modelada por un cuerpo negro calentado solo por su estrella anfitriona, o para planetas fotografiados directamente, la temperatura efectiva del planeta requerida para igualar la luminosidad medida si el planeta fuera un cuerpo negro.

- Temperatura de la estrella modelada por un cuerpo negro que emite la misma cantidad total de radiación electromagnética
- Ascensión Recta del sistema planetario en formato sexagesimal
- Distancia al sistema planetario en unidades de parsec
- Brillo de la estrella anfitriona medido usando la banda V (Johnson) en unidades de magnitudes

2.2 Preprocesamiento de la base de datos

Para implementar algoritmos de ciencia de datos capaces de predecir el tipo de exoplaneta es necesario realizar un preprocesamiento que apoye a tener los datos más relevantes y que mejoren la eficiencia del algoritmo.

En cuanto a la limpieza de datos, es necesario el manejo de valores faltantes para las siguientes variables: Orbit Semi-Major Axis [au], Planet Radius [Earth Radius], Planet Radius [Jupiter Radius], Equilibrium Temperature [K], Stellar Radius [Solar Radius], Stellar Mass [Solar mass], Stellar Surface Gravity [log10(cm/s**2)], Distance [pc] y Gaia Magnitude. Por otro lado, las siguientes variables son removidas debido a la gran cantidad de datos faltantes para la mayoría de las entradas: Planet Mass or Mass*sin(i) [Earth Mass], Planet Mass or Mass*sin(i) [Jupiter Mass], Eccentricity, Insolation Flux [Earth Flux]. Del mismo modo, se removerá la variable Planet Radius [Jupiter Radius], ya que establece la misma longitud desde el centro del planeta hasta su superficie que la variable Planet Radius [Earth Radius] solo que esta utiliza las unidades de radio de Júpiter. En la Tabla 2, se muestra el porcentaje de valores faltantes para cada variable.

Variable	Porcentaje de valor faltante
Eccentricity	88.14
Planet Mass or Mass*sin(i) [Jupiter Mass]	86.72
Planet Mass or Mass*sin(i) [Earth Mass]	86.66
Insolation Flux [Earth Flux]	11.43
Orbit Semi-Major Axis [au]	11.06
Equilibrium Temperature [K]	8.27
Distance [pc]	2.4
Stellar Surface Gravity [log10(cm/s**2)]	1.9
Planet Radius [Jupiter Radius]	1.08
Planet Radius [Earth Radius]	1.01
Gaia Magnitude	0.95
Stellar Mass [Solar mass]	0.74
Stellar Effective Temperature [K]	0.71
Stellar Radius [Solar Radius]	0.31
V (Johnson) Magnitude	0.06
Ks (2MASS) Magnitude	0.06
Planet type	0
Orbital Period [days]	0
Number of planets	0
RA [sexagesimal]	0
Number of stars	0
Planet Name	0

Tabla 2. Porcentaje de datos faltantes.

Fuente: Elaboración propia.

Además, será necesario la transformación de datos, ya que estos manejan rango de error utilizando ± (por ejemplo, 2.6±0.1) por lo que será necesario tomar solo su valor numérico. Se requirió aplicar una expresión regular que separará la cadena de texto de acuerdo a los símbolos '±', '+' o '-', para después tomar el primer valor obtenido.

En el caso de la variable RA [sexagesimal] su valor está en tres unidades: horas, minutos y segundos. Por lo que se convierte a una solo unidad, hora.

Nuevamente, mediante el uso de expresiones regulares se separaron los valores de horas, minutos y segundos, por lo que simplemente al valor de horas se sumó el valor de minutos dividido en 60 y finalmente el de segundos dividido entre 3600.

Una consideración es que el tipo de exoplanetas es una variable categórica, por lo que se transforma utilizando un código numérico:

- Gigante Gaseoso=> 0
- Tipo Neptuno => 1
- Súper Tierra => 2
- Terrestre => 3

Basados en el trabajo de (Rincón Valencia, 2021) se empleó la escala logarítmica para las variables “Orbital Period [days]”, “Orbit Semi-Major Axis [au]”, “Planet Radius [Earth Radius]”, “RA [sexagesimal]” y “Distance [pc]”, debido a que produce una mayor correlación entre las variables.

En la Figura 1 se muestra el histograma de las frecuencias para la variable del Tipo de Exoplaneta; donde son representados los cuatro tipos. Sin embargo, para el caso de Gigante Gaseoso y Terrestre la cantidad de entradas por cada tipo es mucho menor que el caso de Tipo Neptuno y Súper Tierra. Para compensar esta situación se utiliza una técnica de sobremuestreo de minorías sintéticas (SMOTE, del inglés Synthetic Minority Over-sampling Technique), que permite la creación de nuevas entradas a partir del conjunto de datos existente, es decir, equilibra la clase minoritaria frente a la mayorista. Por lo que, después de aplicar el sobremuestreo, cada clase cuenta con 1191 entradas.

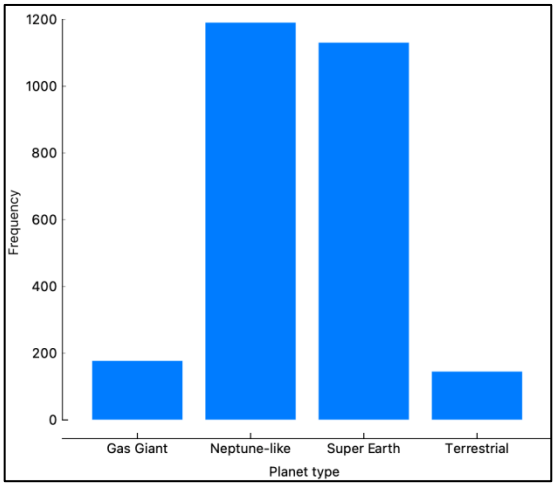

Figura 1. Distribución de la variable Tipo de Exoplanetas.

Fuente: Elaboración propia.

Finalmente, se requiere dividir los datos para el entrenamiento y validación, para esta investigación el 75% de los datos se usarán para el entrenamiento y el 25% para la validación; al iniciar el preprocesamiento la base de datos tenía 3255 exoplanetas y, al terminar, se obtuvieron 4764 exoplanetas con 16 variables; es decir, se usarán 3573 exoplanetas para el entrenamiento y 1191 para la prueba.

2.3 Análisis exploratorio de los datos.

2.3.1. Principales características de la base de datos.

La base de datos cuenta con 4764 entradas después de realizar el preprocesamiento, con 16 variables las cuales se encuentran enlistadas en la Tabla 3, donde la primera columna muestra el nombre de la variable almacenada en el Dataframe y la segunda columna una pequeña descripción del valor de la variable.

Variable	Descripción
Planet type	Los planetas pueden ser clasificados en: Gigante Gaseoso, Súper Tierra, Tipo Neptuno y Terrestre.
Number of stars	Número de estrellas en el sistema planetario.
Number of planets	Número de planetas en el sistema planetario.
Orbital Period [days]	Tiempo que tarda el planeta en realizar una órbita completa alrededor de la estrella o sistema anfitrión
Orbit Semi-Major Axis [au]	El radio más largo de una órbita elíptica o, para los exoplanetas detectados mediante microlente gravitacional o imágenes directas, la separación proyectada en el plano del cielo.
Planet Radius [Earth Radius]	Longitud de un segmento de línea desde el centro del planeta hasta su superficie, medida en unidades de radio de la Tierra
Equilibrium Temperature [K]	La temperatura de equilibrio del planeta modelada por un cuerpo negro calentado solo por su estrella anfitriona, o para planetas fotografiados directamente, la temperatura efectiva del planeta requerida para igualar la luminosidad medida si el planeta fuera un cuerpo negro.
Stellar Effective Temperature [K]	Temperatura de la estrella modelada por un cuerpo negro que emite la misma cantidad total de radiación electromagnética
Stellar Radius [Solar Radius]	Longitud de un segmento de línea desde el centro de la estrella hasta su superficie, medida en unidades de radio del Sol
Stellar Mass [Solar mass]	Cantidad de materia contenida en la estrella, medida en unidades de masa del Sol
Stellar Surface Gravity [log10(cm/s2)]**	Aceleración gravitatoria experimentada en la superficie estelar
RA [sexagesimal]	Ascensión Recta del sistema planetario en formato sexagesimal
Distance [pc]	Distancia al sistema planetario en unidades de parsecs
V (Johnson) Magnitude	Brillo de la estrella anfitriona medido usando la banda V (Johnson) en unidades de magnitudes
Ks (2MASS) Magnitude	Brillo de la estrella anfitriona medido usando la banda K (2MASS) en unidades de magnitudes
Gaia Magnitude	Brillo de la estrella anfitriona medido usando la banda de Gaia en unidades de magnitudes. Objetos emparejados con Gaia utilizando los ID de Hipparcos o 2MASS proporcionados en Gaia DR2.

Tabla 3. Variables en la base de datos después del preprocesamiento.

Fuente: Elaboración propia.

2.3.2. Distribuciones

La variable número de estrellas se compone de enteros donde el valor mínimo es de 1 y el máximo es de 4, con una media de 1.0427 y una desviación estándar de 0.2253. La cantidad más frecuente para esta variable es 1.

La siguiente variable es número de planetas, y al igual que número de estrellas, se trata de un número entero donde el valor mínimo es de 1 y el valor máximo es de 6; con un promedio de 1.8589 y una desviación estándar de 1.1376. En la Figura 2 se muestra el histograma de las frecuencias de la variable Número de Planetas.

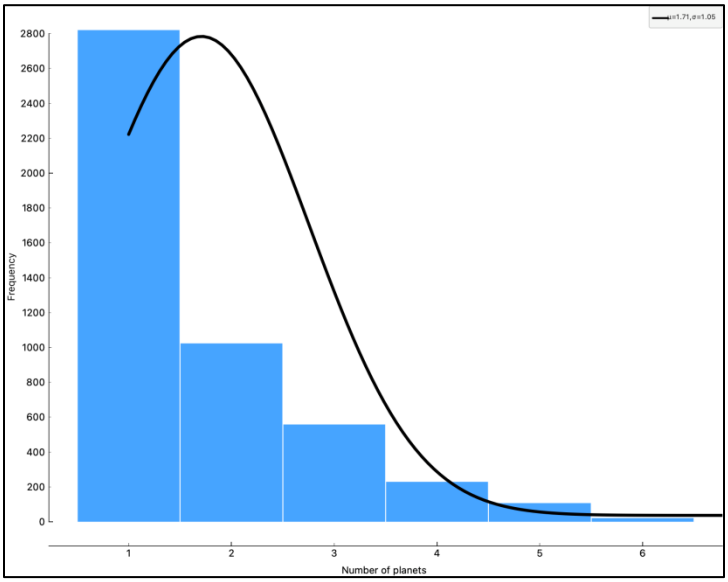

Figura 2. Histograma de la variable Número de Planetas.

Fuente: Elaboración propia.

Para el caso de la variable de periodo orbital está dado en un valor flotante dado en unidades de días. En la Figura 3 se muestra el histograma de la variable periodo orbital.

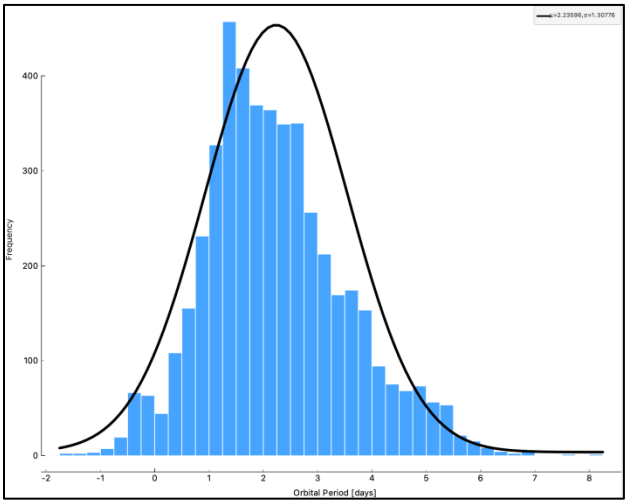

Figura 3. Histograma de frecuencias para variable periodo orbital.

Fuente: Elaboración propia.

La variable órbita semieje mayor está dada en la unidad astronómica (AU), esta unidad es utilizada para medir distancias. Una unidad representa 149,597,871 kilómetros, que es la distancia promedio desde el centro del Sol al centro de la Tierra. El histograma de las frecuencias para esta variable se muestra en la Figura 4.

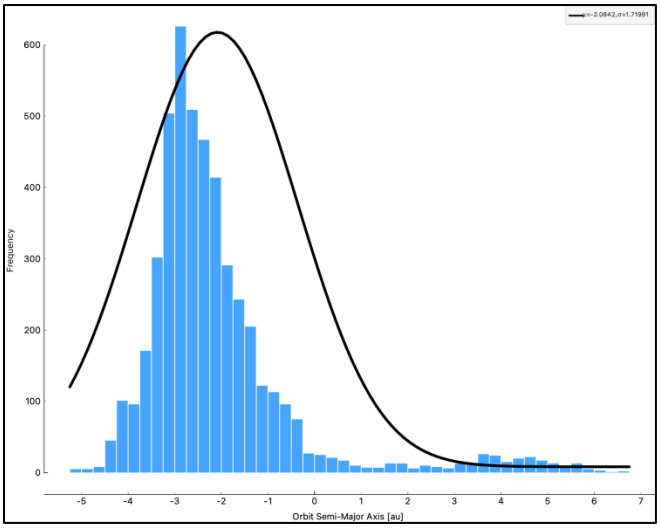

Figura 4. Histograma de frecuencias para la variable Órbita Semieje Mayor.

Fuente: Elaboración propia.

En el caso de la variable Radio del Planeta utiliza la unidad del radio de la Tierra que tiene un valor de 6.371 kilómetros, en la Figura 5 se muestra el histograma de frecuencias para la variable Radio del Planeta.

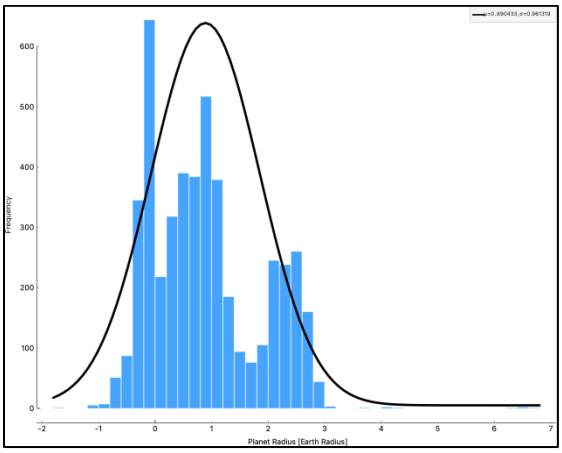

Figura 5. Histograma de frecuencias para la variable Radio del Planeta

Fuente: Elaboración propia.

En el caso de la variable Temperatura de Equilibrio, la unidad es Kelvin. En la Figura 6 se muestra el histograma de las frecuencias para esta variable.

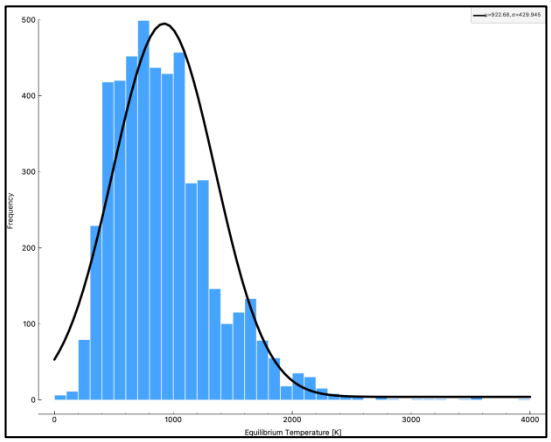

Figura 6. Histograma de la variable Temperatura Equilibrio.

Fuente: Elaboración propia.

La variable Temperatura Efectiva Estelar indica la cantidad de calor que la estrella radia por unidad de superficie, los valores están dados bajo la unidad Kelvin. El histograma de frecuencias se muestra en la Figura 7.

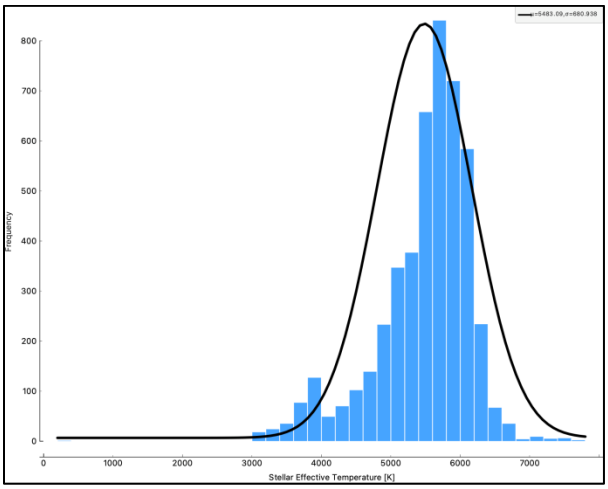

Figura 7. Histograma de frecuencias para la variable Temperatura Efectiva Estelar.

Fuente: Elaboración propia.

La variable Radio Estelar utiliza la unidad del radio del Sol, que equivale a 696,340 kilómetros. La Figura 8 muestra el histograma de las frecuencias para esta variable.

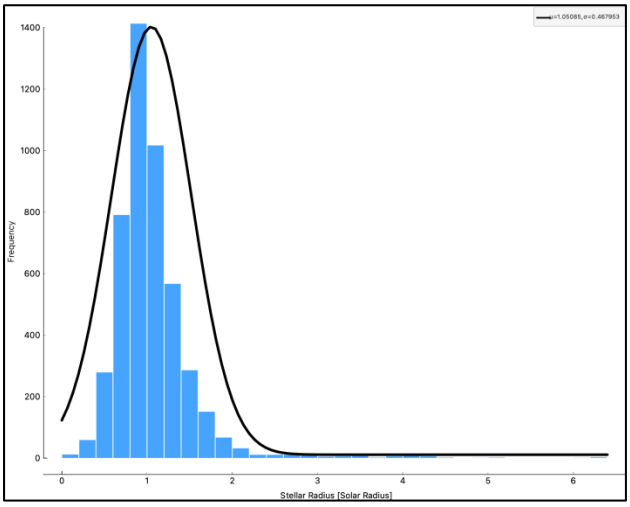

Figura 8. Histograma de la variable Radio Estelar.

Fuente: Elaboración propia.

La variable Masa Estelar, al igual que Radio Estelar, utiliza el Radio del Sol como unidad. En la Figura 9 se muestra el histograma de la variable, con la peculiaridad de ser muy simétrico, es decir, se muestra en el centro el pico más alto y a medida que se aleja del valor central disminuye.

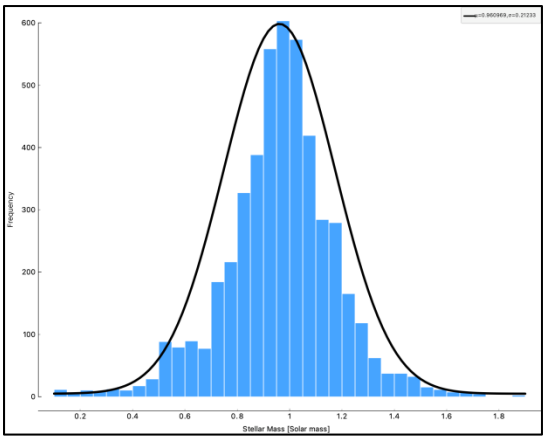

Figura 9. Histograma de la variable Masa Estelar.

Fuente: Elaboración propia.

La variable Gravedad de la Superficie Estelar que es obtenida mediante su tamaño y masa, y gracias a ella se puede inferir etapa evolutiva, masa y otras propiedades físicas. Su histograma de frecuencias se muestra en la Figura 10.

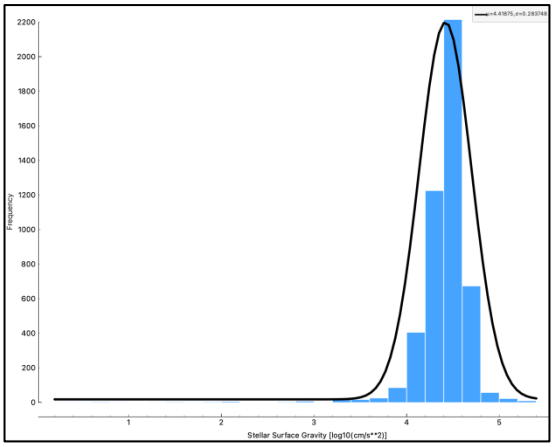

Figura 10. Histograma de la variable Gravedad de la Superficie Estelar

Fuente: Elaboración propia.

Ascensión Recta (RA, por sus siglas en inglés Right Ascension) del sistema planetario en formato sexagesimal, esta variable determina la distancia angular medida hacia el este a lo largo del ecuador celeste desde el CIO, o equinoccio, al círculo horario que pasa por el objeto celeste. La ascensión recta se da ya sea en arco o unidades de tiempo (Capitaine et al., 2007). En la Figura 11 se muestra el histograma de frecuencias para la variable.

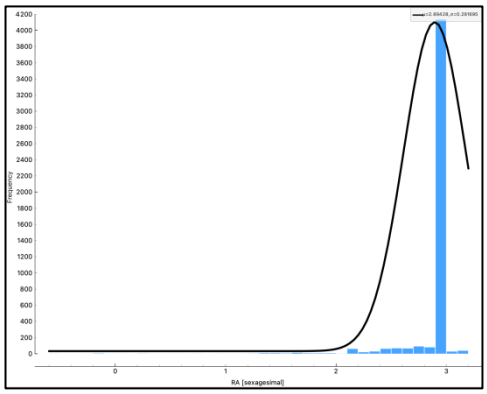

Figura 11. Histograma de la variable Ascensión Recta.

Fuente: Elaboración propia.

Para determinar el brillo de la estrella anfitriona se utilizan dos variables la primera usando la banda V (Johnson) y la segunda usando la banda K (2MASS) ambas dadas en unidades de magnitudes. La Figura 12 muestra el histograma de frecuencias para la variable V (Johnson); por otro lado la Figura 13 muestra el histograma de la variable K (2MASS).

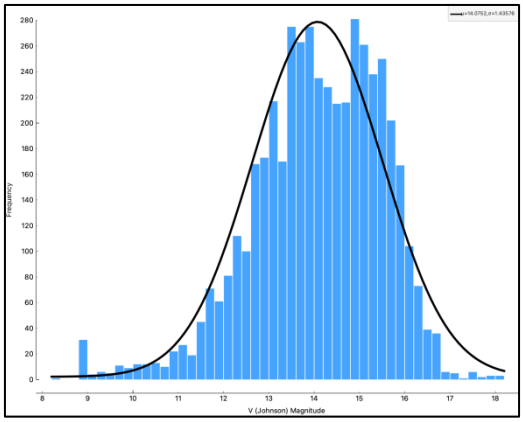

Figura 12. Histograma de la variable V (Johnson)

Fuente: Elaboración propia.

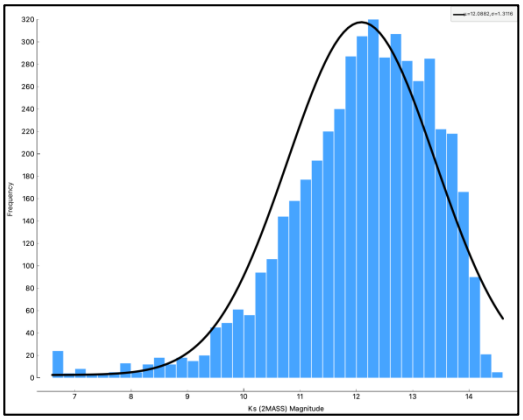

Figura 13. Histograma de frecuencias para la variable K (2MASS).

Fuente: Elaboración propia.

2.3.3. Matriz de Correlación.

Una matriz de correlación permite identificar la calidad de las relaciones entre las variables en un conjunto de datos. Su escala de valores se encuentra dentro del rango [-1,1], donde los valores entre más se acercan a 1 muestran una correlación positiva más fuerte. Los valores iguales a 0 representan una falta de correlación y, a medida que se acercan a –1, representan una correlación negativa más fuerte. La Figura 14 muestra la matriz de correlación utilizando nuestro conjunto de datos.

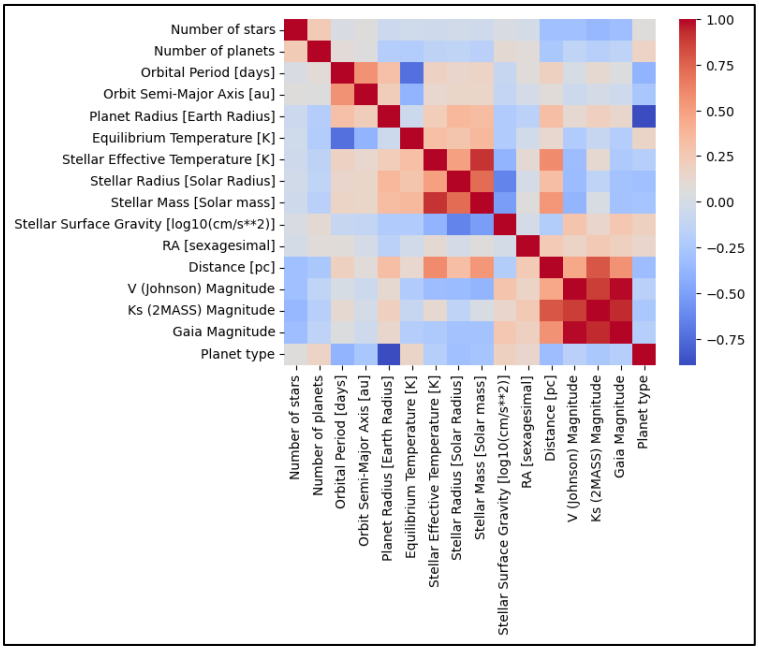

Figura 14. Matriz de Correlación.

Fuente: Elaboración propia.

La diagonal principal de la matriz representa comparaciones consigo mismos; por lo tanto, su valor siempre es 1. En el caso de correlaciones positivas fuertes, como lo son: Stellar Mass con Stellar Effective Temperature y Gaia Magnitude con V (Johnson) Magnitude; lo que nos quiere decir es que a medida que aumente una de estas variables la otra también aumentará y por el contario las correlaciones negativas fuertes como Equilibrium Temperature y Orbital Period los valores descenderán a medida que la otra aumente.

2.3.4. Estandarización.

Es implementada una técnica de normalización conocida estandarización en todas las variables, permitiendo que todas las características tengan una misma escala, haciendo posible una comparación justa entre las características; sin esta normalización el algoritmo puede dar más peso a características con valores más grandes, sesgando el modelo y afectando en el rendimiento. Para aplicar la

normalización, se resta la media y divide por la desviación estándar, haciendo que se tenga una desviación estándar de 1. Es importante mencionar que esta normalización no modifica la distribución de los datos; como ejemplo, en la Figura 15 se muestra la normalización aplicada para la variable “Orbital Period” donde se aprecia que se reduce el rango de (-2,9) a (-3,5) existen variaciones, pero mantiene el comportamiento general de la variable.

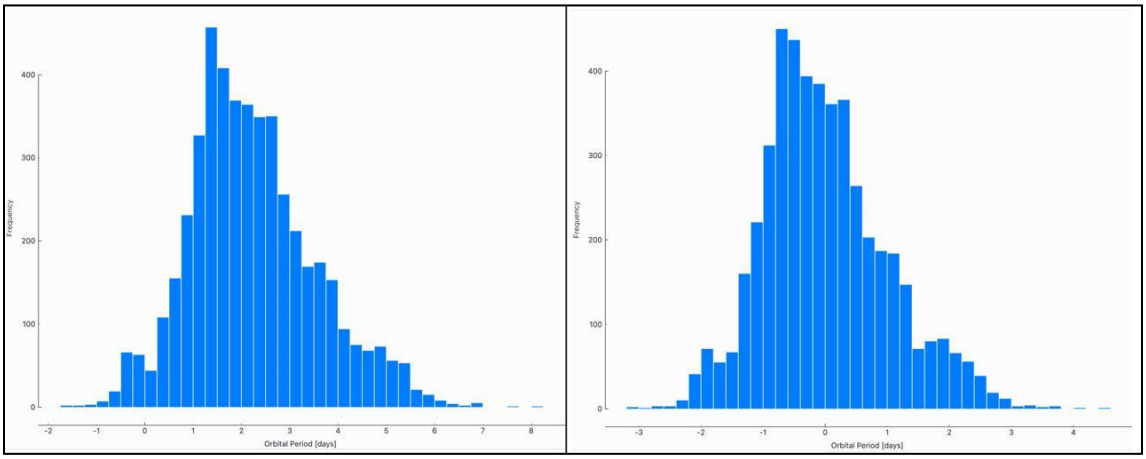

Figura 15. Histogramas que muestran la estandarización para la variable “Orbital Period”.

Fuente: Elaboración propia.

2.3.5. Análisis de Componentes Principales (PCA).

El Análisis de Componentes Principales es una técnica que permite la reducción de dimensionalidad; es decir, reduce el número de variables en el conjunto de datos. Los componentes principales se refieren a las nuevas variables que se derivan de la combinación lineal de las variables originales, donde el primer componente captura la mayor parte de variabilidad en los datos, y así sucesivamente. En la Figura 16 se muestra el gráfico de PCA donde en el eje x muestra la cantidad de componentes principales y el eje y muestra la proporción de varianza, este gráfico es útil para comprender que proporción de varianza en los datos originales es explicada por una componente específica; es decir, cuantos

componentes son necesarios para capturar la mayor parte de información. En el caso de esta investigación, se hará uso de 9 componentes principales.

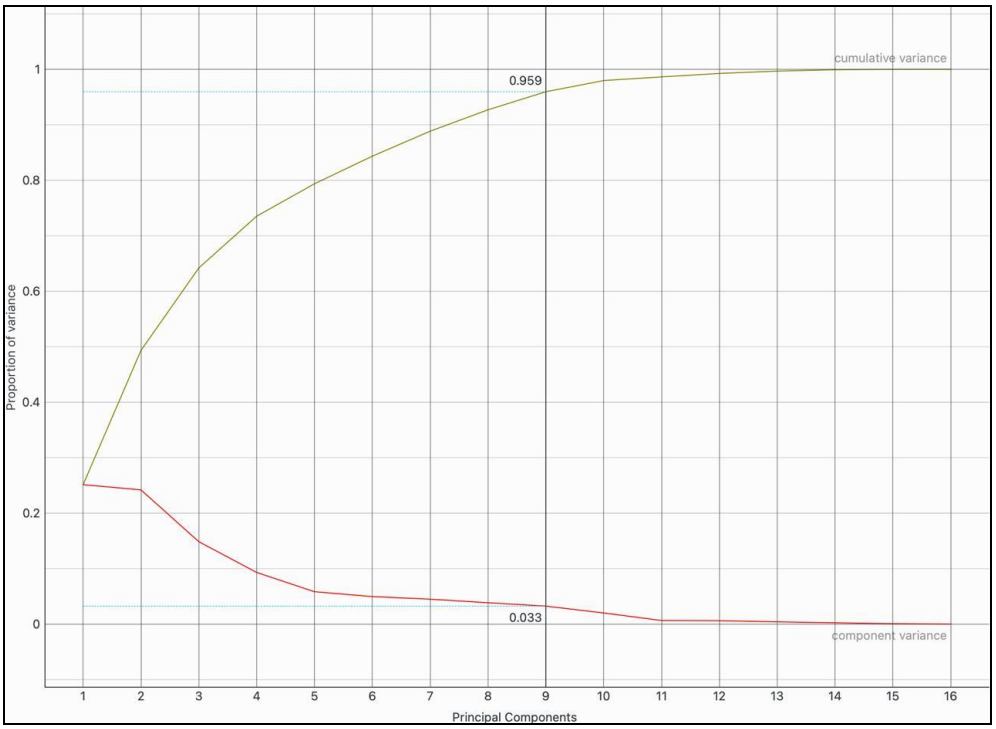

Figura 16. Gráfico de Componentes Principales.

Fuente: Elaboración propia.

Capítulo 3

Diseño del estudio y ajuste de modelos

Capítulo 3. Diseño del estudio y ajuste de modelos

En el siguiente capítulo se tiene como finalidad presentar los conceptos claves utilizados en esta investigación, así como presentar el estado del estudio actual de la clasificación de exoplanetas. Además, de presentar el ajuste de los modelos para esta investigación, que incluye la selección de parámetros y análisis de resultados.

La sección de Marco Teórico expone los conceptos claves para la clasificación de exoplanetas utilizando algoritmos de aprendizaje computacional. Esta sección se encuentra dividida en dos principales áreas de estudio: Astronomía y Ciencia de Datos. A continuación, la sección de Antecedentes provee el contexto de las investigaciones previas o estudios relevantes que sustentan esta investigación.

La sección de Marco Metodológico muestra los pasos metodológicos necesarios para cumplir con los objetivos de esta investigación. La cuarta sección es Ajuste de los Modelos, donde expone el proceso de selección de parámetros y ajuste de los algoritmos. Finalmente, la sección de Análisis de los Resultados muestra la interpretación de los resultados obtenidos de la sección previa.

3.1 Marco teórico.

En la presente sección se abarcan los antecedentes, junto a las principales teorías y los conceptos que dan bases y sustentan la clasificación de los exoplanetas a través de los algoritmos de ciencia de datos.

El marco teórico se divide en dos secciones que representan las principales áreas de estudio en las que se basa esta investigación: Astronomía y Ciencias de Datos.

En la sección de astronomía se introducen conceptos como planeta, exoplanetas, métodos de detección de exoplanetas, misión Kepler y tipos de exoplanetas. Y con respecto a la sección de Ciencia de Datos, se introducen los conceptos de Big Data, aprendizaje de máquina, tipos de aprendizaje y Redes Neuronales.

3.1.1 Astronomía.

La palabra astronomía, proviene del griego antiguo “astron” que significa estrellas y “nomos” que significa ley o norma. Es la ciencia que estudia los cuerpos celestes que pueblan el cosmos. Los cuerpos celestes o astros son las masas de materia que constituye el Universo (El Universo, 2013). Algunos ejemplos son los planetas, estrellas, satélites, meteoritos y galaxias.

“Planeta” proviene del griego errante o vagabundo, es un astro opaco redondo que gira alrededor de una estrella y que domina su entorno, ya que su masa es suficientemente grande como para atraer la mayor parte de la materia que se aproxime a él (El Universo, 2013); por opaco, se refiere a que es un astro que no emite luz propia, solamente refleja la que le llega (García, 2018) esto se debe a que los planetas no tienen suficiente masa para alcanzar en sus interiores las temperaturas necesarias para generar energía por reacciones nucleares como hacen las estrellas, lo que hace difícil su detección directa (Rebolo, 2003).

Cuando los planetas orbitan estrellas diferentes a la nuestra, conocido como Sol, se les denomina exoplanetas o planetas extrasolares (Ruíz, 2017). En 1995, los suizos Michel Mayor y Didier Queloz publicaron un descubrimiento histórico: la detección del primer exoplaneta (Hernández, 2003), su composición y estructura se infiere de datos en conjunto con otras medidas indirectas como la espectrometría, induciendo a partir de esos datos parciales y fragmentados su historia geológica y características ambientales (Ruíz, 2017). Desde 2015 hasta

hoy, el número de exoplanetas se ha multiplicado por más de 2: de 2000 a casi 4500. Lo ideal para detectar exoplanetas sería poder verlos directamente, sin valerse de métodos indirectos, sino solo resolviendo la luz de las dos o tres componentes, estrella y planeta o planeta + satélite, para visualizarlas en el telescopio (Monzón, 2021). Algunos otros métodos de detección son:

- El método Doppler de velocidad radial tiene en cuenta el vaivén gravitatorio que los exoplanetas gigantes muy masivos provocan en sus estrellas. El vaivén será mayor cuanto más masivo sea el exoplaneta respecto a su estrella y cuanto más cerca esté de ella (Hernández, 2003).
- El método astrométrico se basa en el mismo efecto que el de la velocidad radial, pero en este no se mide el desplazamiento, sino que se mide directamente la distancia angular que hay entre el baricentro del sistema solar y el centro de la estrella (Hernández, 2003).
- En el caso de los cuerpos de masa planetaria sin ligadura gravitatoria a estrellas, la detección se ha realizado de manera directa y se dispone tanto de imágenes como de espectros que han permitido determinar sus propiedades físicas básicas (Rebolo, 2003).
- El método de Tránsito usa la posibilidad de ser detectado mediante fotometría. Se basa en la caída de luz que apreciamos cuando un cuerpo se interpone entre una estrella y nuestra línea de visión, lo que provoca que el disco de luz de la estrella se oscurezca mínimamente debido al tránsito del planeta por delante del disco estelar. La radiación que captamos del exoplaneta proviene tanto de la luz que refleja de la propia estrella como del calor que despide. La luz reflejada por el planeta, por desgracia, suele ser muy pequeña en comparación con la que emite la estrella (Monzón, 2021).

Los métodos de velocidad radial y tránsitos, que han sido responsables del descubrimiento del 96% del total de exoplanetas (Monzón, 2021).

La misión Kepler de la NASA es considerada una de las más emblemáticas para buscar exoplanetas. Comenzó en el año 2009 cuando se puso en órbita. Algunas de sus características son: espejo primario de 1.4 metros, un fotómetro

Schmidt de 0.95 metros de apertura y una cámara CCD de 95 millones de píxeles de resolución. Este equipo permitió monitorizar a más de 150,000 estrellas en la constelación de Cisne para la búsqueda de exoplanetas, donde se utilizó el método del tránsito para su detección. Se tenía planeado que la misión duraría 3 años y medio. En el año 2013 fallaron los giroscopios afectando el sistema de orientación del equipo; en ese momento ya se habían detectado y confirmado más de 2330 exoplanetas (García, 2018). Con el lanzamiento del telescopio Kepler en 2010, se produjo un punto de inflexión en la historia del descubrimiento de exoplanetas, debido en parte a su potentísima precisión fotométrica, pudiendo llegar a detectar un cambio en el flujo estelar de la estrella de ~0.0023 milésimas de magnitud y al amplio seguimiento de 150000 estrellas. Las misiones Kepler y K2 han sido exitosas, habiendo descubierto 2347 exoplanetas confirmados y 2420 candidatos mediante el método de tránsito. En 2018, estas misiones llegaron a su fin (Monzón, 2021).

3.1.1.1 Tipos de exoplanetas.

Hasta ahora los científicos han clasificado los exoplanetas en los siguientes tipos: Gigante Gaseoso, Neptuniano, Súper Tierra y Terrestre, donde el tamaño y la masa juegan un papel crucial en la determinación de los tipos (Overview | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2022). Sin embargo, se determinó gracias a la misión Kepler que los planetas que son 1.5 y 2 veces mayores al diámetro de la Tierra son raros.

3.1.1.1.1 Gigante Gaseoso

Los exoplanetas Gigantes Gaseosos son esferas gigantes compuestas de hidrógeno y/o helio, aunque no se sabe si todos los gigantes gaseosos tienen núcleo. Se ha observado que tienen un mecanismo de inflación interior que no permite que se encogiera según los modelos de enfriamiento. El tener núcleo pesado en su interior permitiría conocer y posiblemente explicar el comportamiento (Agudelo, 2022).

En nuestro Sistema Solar los planetas Júpiter y Saturno son catalogados como Gigantes Gaseosos, en general no presentan superficies duras, sino que tienen gases arremolinados sobre un núcleo sólido; estos generalmente se encuentran orbitando extremadamente cerca de sus estrellas madre, dándoles vuelta en 18 horas y se cree que se forman dentro de los primeros 10 millones de años de vida de una estrella similar al Sol o no se forman en absoluto (Gas Giant | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2022).

Debido a las técnicas de detección de exoplanetas, la mayoría de los exoplanetas encontrados son gigantes gaseosos similares a Júpiter, pero con órbitas más cercanas a sus estrellas que Mercurio al Sol (Curiel & Curiel Ramírez, 2011), ya que estos grandes planetas forman órbitas tan estrechas que provocan una pronunciada “bamboleo” en sus estrellas, empujando sus estrellas anfitrionas de un lado a otro, y provocando un cambio mensurable en el espectro de luz de las estrellas (Gas Giant | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2022).

3.1.1.1.2 Neptuniano.

Los exoplanetas tipo Neptuno poseen un tamaño similar a Neptuno o Urano, con aproximadamente cuatro veces el tamaño o radio de la Tierra y casi 17 veces su masa o peso (Neptune-like | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2023) se consideran planetas con un grueso envoltorio (Agudelo, 2022), es decir sus atmósferas suelen estar dominadas por hidrógeno y helio (aunque es común que nubes espesas bloquen el paso de la luz, evitando la visibilidad de las moléculas en la atmósfera) con núcleos de roca y metales pesados (Neptune-like | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2023), compuestos en su gran mayoría de masa agua, amonio y metano; tienen un recubrimiento a base de agua y corteza terrestre y no todo el planeta se encuentra condensado (Agudelo, 2022).

En 2016, un estudio encontró que los mundos con la masa de Neptuno pueden ser el tipo de planeta más común, aproximadamente 10 veces más comunes que los planetas con la masa de Júpiter, que se forma en los reinos exteriores helados de los sistemas planetarios (Neptune-like | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2023).

3.1.1.1.3 Súper Tierra.

Las Súper Tierras son consideradas más masivas que la Tierra, pero más ligeras que los gigantes de hielo como lo son Neptuno y Urano. Estos planetas pueden estar hechos de gas, roca o una combinación de ambos. Su tamaño puede ser el doble que el de la Tierra y tener hasta 10 veces su masa. En nuestro sistema solar no tenemos esta clase de planetas. Su nombre solo hace referencia al tamaño del planeta, no establece una necesaria similitud con nuestro planeta Tierra. Se cree que podría existir una amplia variedad de composiciones planetarias, como lo son los mundos acuáticos, planetas bola de nieve o compuestos principalmente de gas denso (Super-Earth | Planet Types – Exoplanet Exploration: Planets beyond our Solar System, 2022).

3.1.1.1.4 Terrestre.

En el sistema solar se cuentan con un gran número de ejemplos del tipo Terrestre y estos son la Tierra, Marte, Mercurio y Venus. También se le puede llamar planeta rocoso. Son considerados terrestres aquellos planetas que midan entre la mitad y el doble del radio de la Tierra, aunque pueden existir algunos otros que pueden llegar a ser incluso más pequeños. Aunque pueden existir exoplanetas rocosos del doble del tamaño de la Tierra, estos son considerados Súper Tierras. Se les llama planetas terrestres, ya que se les considera mundos rocosos; es decir, se encuentran compuestos de roca, silicato, agua y/o carbono; con una superficie sólida o líquida. Existen menos mundos rocosos y potencialmente

habitables entre los miles de exoplanetas encontrados hasta ahora (Terrestrial | Planet Types – Exoplanet Exploration: Planets beyond our solar System, 2022).

La habitabilidad se definirá como la capacidad que tiene un exoplaneta de sostener la vida y esta se asocia con la presencia de agua líquida en su superficie. Y una zona habitable como una cierta restricción sobre la órbita de un planeta con respecto a su estrella, la cual permite al planeta mantener agua líquida en su superficie (Monzón, 2021).

3.1.2 Ciencia de Datos.

El Descubrimiento de Conocimiento en Bases de Datos (KDD, del inglés Knowledge Discovery in Databases) busca analizar grandes cantidades de datos, encontrando relaciones o patrones que puedan generar información. Algunas herramientas para lograrlo son el aprendizaje automático, aplicación de estadística, implementación de base de datos, representación del conocimiento, razonamiento basado en casos y aproximado, adquisición de conocimiento, redes neuronales y la visualización de los datos (García, 2018).

Según un informe de IBM, el 90% de los datos disponibles actualmente en el mundo se han creado en los dos últimos años (García, 2018). De manera más específica, hoy en día la cantidad de datos astronómicos (tanto observacionales como teóricos) en archivos sobrepasan los Petabytes (Molina, 2021). Por lo que a las cantidades masivas de datos recogidas a lo largo del tiempo responden al concepto de Big Data, o datos masivos; donde los conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenamiento, gestión y análisis de las herramientas de base de datos (Maté, 2014). Es importante aclarar que no solo se hace referencia al tamaño de la información, sino también a la variedad del contenido y a la velocidad con la que los datos se generan, almacenan y analizan (García, 2018).

En esta nueva era no solo es necesario contar con grandes infraestructuras de almacenamiento y altas velocidades de transferencia, sino también que es prioritario desarrollar herramientas de visualización que posibiliten realizar análisis de manera ágil y rápida (Molina, 2021). Siendo la ciencia de datos un campo interdisciplinario, que utiliza Machine Learning, Estadística, Minería de Datos y Programación, para la extracción de conocimiento que puede parecer oculto a partir de datos, se utilizan procesos de descubrimiento, formulación y verificación de hipótesis (García, 2018).

Aprender es la habilidad de adquirir conocimientos, desarrollar habilidades para analizar y evaluar problemas mediante métodos y técnicas utilizando la experiencia propia; el aprendizaje en los sistemas puede definirse como el cambio que experimenta permitiéndole resolver mejor una tarea la segunda vez u otra tarea similar (Chaviano Arteaga, 2015).

3.1.2.1 Tipos de Aprendizaje

3.1.2.1.1 Supervisado

El Aprendizaje Supervisado constituye un algoritmo de aprendizaje basado en ejemplos donde el nuevo conocimiento es inducido a partir de una serie de ejemplos y contraejemplos, por lo tanto, existe una correspondencia entre las entradas y salidas deseadas del sistema (Chaviano Arteaga, 2015).

La clasificación y regresión son los dos tipos de problemas que pueden emplear el aprendizaje supervisado. La tarea de clasificación consiste en encontrar una función objetivo capaz de estimar la clase correcta para un objeto. Dado el conjunto de entrenamiento, se pretende identificar el mejor clasificador en el espacio de hipótesis, utilizando un algoritmo de aprendizaje que realiza una búsqueda en el espacio tomando como base el conjunto de entrenamiento, infiriendo así la función de clasificación. (Águila, 2017).

3.1.2.1.1.1 Redes Neuronales Artificiales.

Las Redes Neuronales Artificiales se empezaron a investigar en 1930, donde su capacidad de detección y aprendizaje de patrones en un conjunto de datos generaron una nueva herramienta para la resolución de problemas que requieren el análisis de información (García, 2018). Se basan en la analogía del comportamiento y función del cerebro humano, en particular del sistema nervioso, compuesto por redes de neuronas biológicas con bajas capacidades de procesamiento, pero su capacidad cognitiva se sustenta en la conectividad de estas (Salas, 2016).

Las Redes Neuronales Artificiales se componen de un conjunto de elementos de proceso, denominadas neuronas, localizadas en los vértices de un grafo dirigido en una estructura que permite la propagación de información de neuronas anteriores hacia neuronas posteriores: cada elemento recibe una entrada (señal) de las unidades anteriores y comunica su salida a las unidades posteriores (Águila, 2017). Las neuronas se encuentran conectadas por dendritas y cada una de ellas posee un valor distinto, denominado peso. Estos tienen un gran rol dentro de la red porque fortalece o debilita la comunicación o conexión entre las neuronas (García, 2018). Tomando como base el diagrama de entrenamiento y aprendizaje propuesto por (Trujillo & Cuevas, 2006), se generó el diagrama mostrado en la Figura 17, donde se plantea de una manera más general el proceso de entrenamiento y aprendizaje para las redes neuronales.

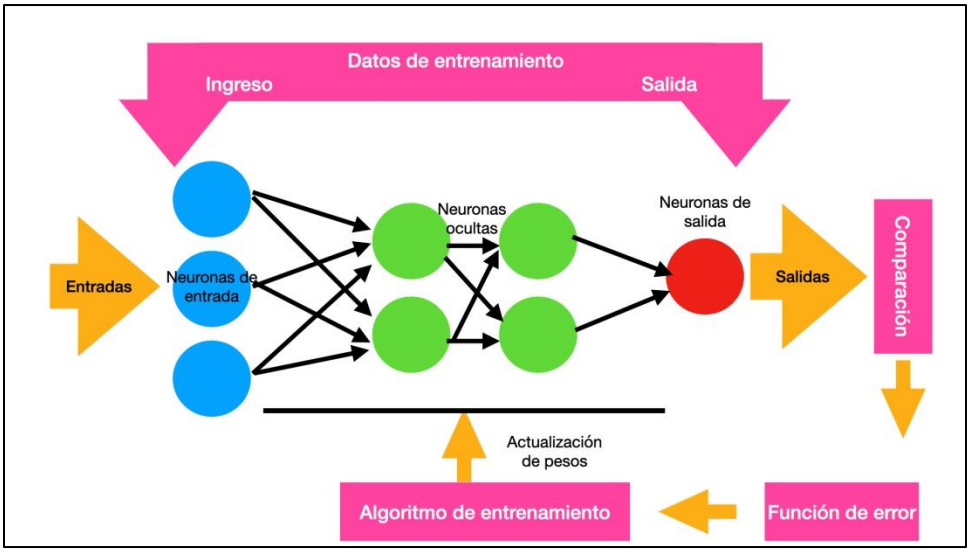

Figura 17. Diagrama de entrenamiento y aprendizaje de una red neuronal.

Fuente: Trujillo & Cuevas, 2006.

Las Redes Neuronales aprenden durante el proceso de entrenamiento, siendo este una parte importante del algoritmo, por eso algunos autores establecen que es la etapa que caracteriza a este conjunto de técnicas. El modelo está compuesto por pesos equivalentes a las conexiones sinápticas, con un umbral de acción o activación (Salas, 2016). Se inicializan aleatoriamente y, a medida que avance el entrenamiento, estas son modificadas. Cuando la red termina el proceso de entrenamiento, es capaz de realizar predicciones, clasificaciones y segmentaciones (García, 2018). Este proceso se puede visualizar en la Figura 18, que es obtenida del trabajo de (Castro Becerra, 2016), donde se muestra el Algoritmo de Entrenamiento para una Red Neuronal, con un criterio de paro dado un número de experimentos.

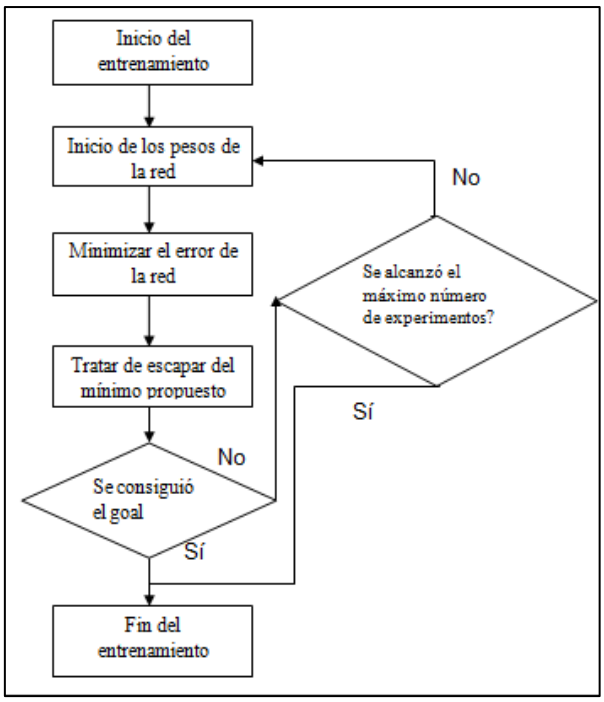

Figura 18. Algoritmo de Entrenamiento para una Red Neuronal.

Fuente: Castro Becerra, 2016.

Las tareas de clasificación supervisada con Redes Neuronales Artificiales consisten en presentarle a la red repetidamente datos de entrada consistentes en un patrón de estímulos y la respuesta esperada, para comparar dicha respuesta con la respuesta que da la red. En virtud a esa comparación, se realizan ajustes increméntales del modelo para que el resultado previsto se acerque cada vez más a la respuesta esperada (Águila, 2017). Algunos trabajos como el de (García, 2018) determinan que la cantidad de neuronas ocultas se determina por un proceso de prueba y error, es decir, a medida que el entrenamiento avance, se irán probando diferentes valores y se quedarán con aquel que muestre mejores resultados. Por otro lado, también establece que el criterio de parada se obtiene al sumar los errores en el patrón de entrenamiento y cuando el error se convierta en constante con la iteración anterior, se mantienen los parámetros y se calcula el error mínimo (García, 2018).

3.1.2.1.1.1 Árboles de decisión.

Los árboles de decisión son modelos no paramétricos de aprendizaje supervisado, cuyo objetivo es poder predecir a qué clase pertenece un caso del que conocemos uno o más atributos o mediciones (Arana, 2021). El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver (Martínez et al., 2009). La Figura 19 muestra la estructura y componentes de un árbol de decisión.

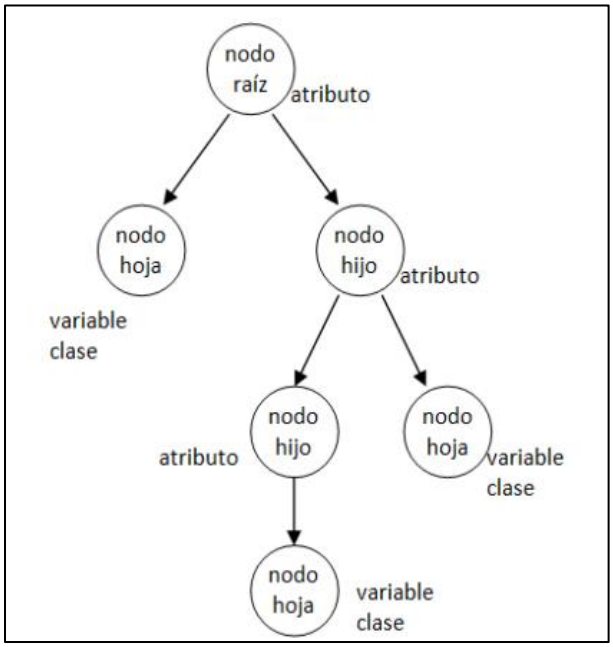

Figura 19. Estructura de un árbol de decisión.

Fuente: Martínez et al., 2009.

3.1.2.1.1.1 Regresión Logística.

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo, pero de manera general formula un análisis utilizado para predecir una variable categórica en función de una o varias variables predictoras. El análisis de regresión logística está inmersa en el grupo que usa una función de enlace llamada logit (Rodríguez & Gallardo, 2020) que correlaciona la probabilidad de una variable cualitativa con un conjunto de variables escalares.

3.1.2.1.1.1 Máquinas de Soporte Vectorial.

Las Máquinas de Soporte Vectorial buscan generar un hiperplano que permita separar una clase de otra, maximizando la distancia entre los puntos de diferentes clases y una función separadora. Un SVM no lineal usa varias funciones kernel para estimar el margen. Algunos ejemplos de estas funciones kernel son: lineal, polinomial, de base radial y sigmoidal (Galindo et al., 2020). Las SVM están ganando popularidad como herramienta para la identificación de sistemas no lineales, esto es a causa, principalmente, a que las SVM están basadas en el principio de minimización del riesgo estructural que permite escoger un clasificador que minimiza una cota superior sobre el riesgo, y proporciona una buena medida para obtener clasificadores que generalizan bien sobre datos no previamente vistos (Alfaro, 2010).

3.1.2.1.2 No Supervisado.

En el Aprendizaje No Supervisado se modela el proceso sobre un conjunto de ejemplos formado solo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos (Chaviano Arteaga, 2015).

Las siguientes técnicas son utilizadas por el Aprendizaje No Supervisado: Clustering, Reducción de Dimensionalidad y Detección de Anomalías. El

Clustering busca formar grupos a partir de un conjunto de datos a partir de las similitudes con otros miembros; por lo tanto, si dos datos pertenecen a un mismo grupo, estos presentarán muchas similitudes, y si los dos datos pertenecen a grupos separados, estos se diferenciarán lo más posible. Por lo que el proceso se centra en lugar de analizar los datos etiquetados en una clase; busca analizar para generar una etiqueta. Al principio se desconoce cuántos grupos de clasificación existirán ni cómo serán estos grupos. Aquí se barajan dos opciones: o “forzar” el algoritmo para obtener un número determinado de grupos o dejar que la herramienta analice la naturaleza de los datos y calcule cuántos grupos sería deseable obtener (García, 2018).

3.1.2.1.2.1 Algoritmo de K-means.

K-means es un algoritmo de clusterización, donde representa cada clúster por la media de sus puntos, generando un centroide, obteniendo un significado gráfico y estadístico inmediato. La suma de las discrepancias entre un punto y su centroide, expresado a través de la distancia apropiada, se usa como función objetivo (Garre et al., 2007). El método de k medias es un procedimiento muy útil a la hora de buscar agrupaciones en numerosos conjuntos de datos, pero debido a sus características intrínsecas presenta ciertas deficiencias y debilidades con determinados conjuntos de datos; por ejemplo, es un método muy sensible ante contaminaciones o pequeñas desviaciones del modelo, ya que si existen puntos aislados del resto, su aportación al potencial sería muy grande y para disminuirlo, el método de k medias tendería a desplazar sus centros en esa dirección a fin de reducirlo (Carlos & De Valladolid Facultad de Ciencias, 2022). En la Figura 20 se encuentra el algoritmo obtenido del trabajo de (Bustamam et al., 2017)

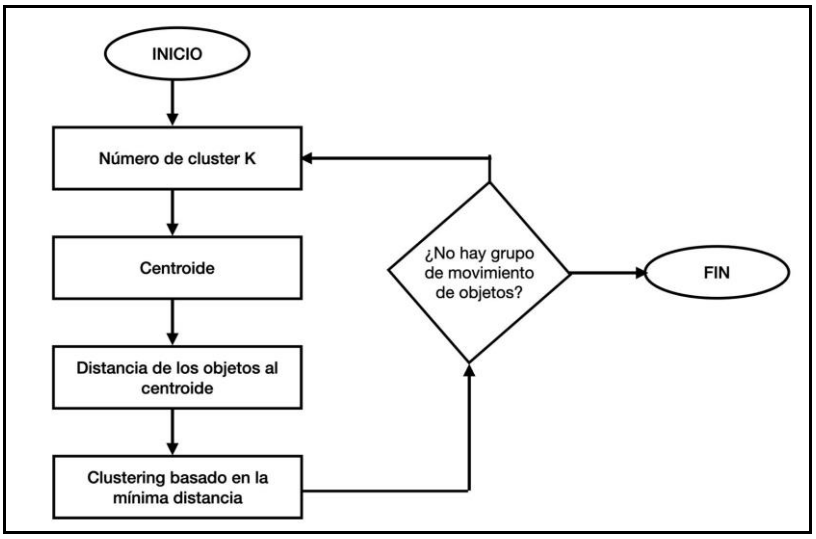

Figura 20. Algoritmo K-means.

Fuente: Bustamam et al., 2017.

3.2 Antecedentes

En las dos últimas décadas hemos experimentado una revolución que nos ha llevado a descubrir miles de nuevos planetas (Ribas, 2019); lo que ha generado una necesidad de clasificación y caracterización en grupos y subgrupos para ello se requiere trazar relaciones, agrupar, categorizar y extraer información a partir de la gran cantidad de datos generados por estos descubrimientos (Chinellato et al., 2022).

Para procesar esta gran cantidad de información se recurrió al aprendizaje computacional; por ejemplo, el trabajo de (Del Hierro Diez, 2020), que utiliza datos con series temporales de intensidad de luz recogida de 5087 estrellas, en 3198 momentos. Cada una de las observaciones o estrellas está etiquetada con “2” o “1” si dicha estrella poseía un exoplaneta orbitándola o no, respectivamente. Se implementan dos variantes de un mapa autoorganizado con diferentes funciones de vecindad; donde solo se actualizarán las neuronas dentro de un radio rectangular alrededor de la neurona ganadora. La Tabla 4 es obtenida del trabajo de (Del Hierro Diez, 2020), donde se muestra la precisión de los mapas

autoorganizados, donde encuentra que la segunda variante con vecindad gaussiana pareció conseguir mejores clasificaciones.

Tamaño mapa	Tratamiento Datos	Precisión Entrenamiento	Precisión Prueba
8x8	-	0.98	0.98
	SMOTE	0.79	0.98
	FT	0.99	0.98
	Sub-muestreo	0.83	0.97
16x16	-	0.99	0.99
	SMOTE	0.8	0.99
	FT	0.99	0.99
	Sub-muestreo	0.51	0.99

Tabla 4. Precisión de algoritmos basados en Mapas Autoorganizados.

Fuente: Elaboración propia con información extraída de (Del Hierro Diez, 2020).

A continuación, se enlistan trabajos donde su principal objetivo es la clasificación de tipo de exoplanetas.

Un primer trabajo es (Chinellato et al., 2022), donde implementa dos algoritmos de aprendizaje automático, K-means que representa al aprendizaje no supervisado y un Árbol de Decisión para el aprendizaje supervisado. El trabajo abarca desde el preprocesamiento de datos hasta la clasificación en grupos, donde la clasificación se da en dos etapas; la primera clasifica en los siguientes grupos: Exotierras, Exojúpiters y planetas de Transición; y la segunda etapa divide la categoría de planetas de Transición en Júpiter Densos y Gigantes de Hielo. La Tabla 5 es extraída del artículo de Chinellato donde muestra los resultados de la clasificación usando k-means; además, en el caso de su implementación con árboles de decisión, logra obtener una entropía nula en cada nodo, es decir, se logra discriminar correctamente cada grupo.

Nombres propuestos (cluster)	Cantidad de miembros	Densidad promedio [g/cm^3]	Radio promedio [R_{Jup}]
Exotierra (cluster 0)	98	2±1	0.25±0.09
Exojúpiter (cluster 1)	182	0.4±0.2	1.2±0.4
Planeta de transición (cluster 2)	156	1.6±0.9	1.1±0.3

Tabla 5. Resultados de clasificación con k-means.

Fuente: Chinellato et al., 2022.

El trabajo de (Meléndez Lorenzo, 2022), por otro lado, utiliza los datos obtenidos por la plataforma Kaggle, reuniendo series temporales de flujo de luz recibido desde diferentes estrellas observadas. Se reúnen, 5087 estrellas para cada una de las cuales se tomaron 3198 medidas de flujo de luz. Del total de estrellas, solo 37 están confirmadas con presencia de exoplanetas. Los métodos implementados fueron: Regresión Logística, Máquinas de Soporte Vectorial, Árboles de decisión, K vecinos, Random Forests y la combinación de modelos por votación. Las precisiones de los algoritmos se encuentran en la Tabla 6. Aunque Meléndez establece que la cantidad de instancias positivas es pequeña para establecer cuál modelo funciona mejor, también afirma que con los resultados actuales el modelo Random Forest presenta una mayor puntuación.

Método	Regresión Logística	Máquinas de Soporte Vectorial	Árboles de decisión	K vecinos	Random Forests	Combinación de modelos por votación
Precisión	0.57±0.13	Kernel lineal 0.97±0.05	0.97±0.05	0.01±0.01	1.0	1.0
		Kernel RTF 0.45±0.12				
		Kernel polinomial 0.53±0.14				

Tabla 6. Precisión de algoritmos utilizando series temporales de flujo de luz.

Fuente: Elaboración propia con información extraída de (Meléndez Lorenzo, 2022).

Por otro lado, el trabajo de (González Cangrejo, 2021) implementa los siguientes algoritmos de aprendizaje supervisado: Árboles de Decisiones, K vecinos más cercanos y Máquinas de Vectores de Soporte; además, utilizando la base de datos de la NASA toma los siguientes atributos: masa del planeta, radio del planeta, periodo orbital y eje semimayor. En este trabajo los algoritmos son modelados cuatro veces: en la primera se implementan masa del planeta y radio del planeta, en la segunda periodo orbital y eje semimayor, posteriormente son sustituidos por masa del planeta, radio del planeta y período orbital, y como último modelo se implementan todas las características. En la Tabla 7 se muestran las precisiones por cada algoritmo; donde se logra corroborar que las técnicas de aprendizaje supervisado de Máquinas de Vectores y el algoritmo Árbol de Decisiones son considerados modelos eficaces en tareas de clasificación y regresión.

Método	Máquinas de Vectores	K Vecinos	Árbol de decisión.
Precisión	0.9552	0.9492	0.973134

Tabla 7. Precisión de los algoritmos: Máquinas de Vectores, K Vecinos y Árbol de decisión.

Fuente: Elaboración propia con información extraída de (González Cangrejo, 2021).

El trabajo de (Rincón Valencia, 2021) clasifica los exoplanetas en Súper Tierra, Gigante Gaseoso, Tipo Neptuno y Terrestre. En esta ocasión utiliza la base de datos de la NASA, pero utiliza las misiones: Tess, Kepler, K2, KELT y Ukirt. Además, en comparación con el trabajo de (González Cangrejo, 2021), utiliza las siguientes variables: nombre del planeta, masa del planeta con respecto a la tierra, radio del planeta con respecto a la tierra, periodo orbital y tipo de planeta. Otra diferencia, es el uso de aprendizaje no supervisado con los algoritmos de K means y clustering jerárquico, que además implementa regresión logística, K vecinos y árbol de decisiones. En la Tabla 8 se muestran las precisiones por cada algoritmo. En los algoritmos de Aprendizaje No Supervisado no logra asegurar el comportamiento del clasificador, ya que lo considera un método menos preciso y

fiable, ya que no logra información precisa sobre la clasificación de los datos y la salida.

Método	Regresión Logística	K Vecinos	Árbol de decisión.
Precisión	0.919453	0.9346	0.9483

Tabla 8. Precisión de los algoritmos: Regresión Logística, K Vecinos y Árbol de decisión.

Fuente: Elaboración propia con información extraída (Rincón Valencia, 2021).

En conclusión, gracias a las misiones espaciales se han generado una gran cantidad de datos, que han sido aprovechados por diferentes proyectos o investigaciones. Podemos encontrar trabajos que utilizan el Archivo de exoplanetas de la NASA, que es una de las fuentes de datos para esta investigación, pero lo hacen para su detección, o algunos otros trabajos que también clasifican el tipo de exoplaneta, como lo es el trabajo de (González Cangrejo, 2021) que implementa los siguientes algoritmos de aprendizaje supervisado: Árboles de Decisión, K-vecinos y Máquinas de Soporte Vectorial utilizando las variables: masa del planeta, radio del planeta, periodo orbital y eje semimayor; o el trabajo de (Rincón Valencia, 2021) que implementa tanto algoritmos de aprendizaje supervisado (Árboles Decisión, K-vecinos y Regresión logística) como no supervisado (K media) utilizando tres variables: tamaño, masa del planeta y periodo orbital. Este trabajo se diferencia de los anteriores por la cantidad de variables, ya que se utilizan 16, y se implementarán Redes Neuronales como el algoritmo de clasificación.

3.3 Marco metodológico

La presente investigación posee un enfoque cuantitativo con un alcance exploratorio; por lo tanto, se busca clasificar desde una nueva perspectiva los tipos de exoplanetas utilizando variables del sistema planetario. Los pasos metodológicos de esta investigación se plasman en la Figura 21, divididos en 4

fases principales donde responden a un objetivo específico; y al completar estas fases se logra el objetivo de esta investigación. En las siguientes secciones se describen las fases con sus pasos específicos para su desarrollo.

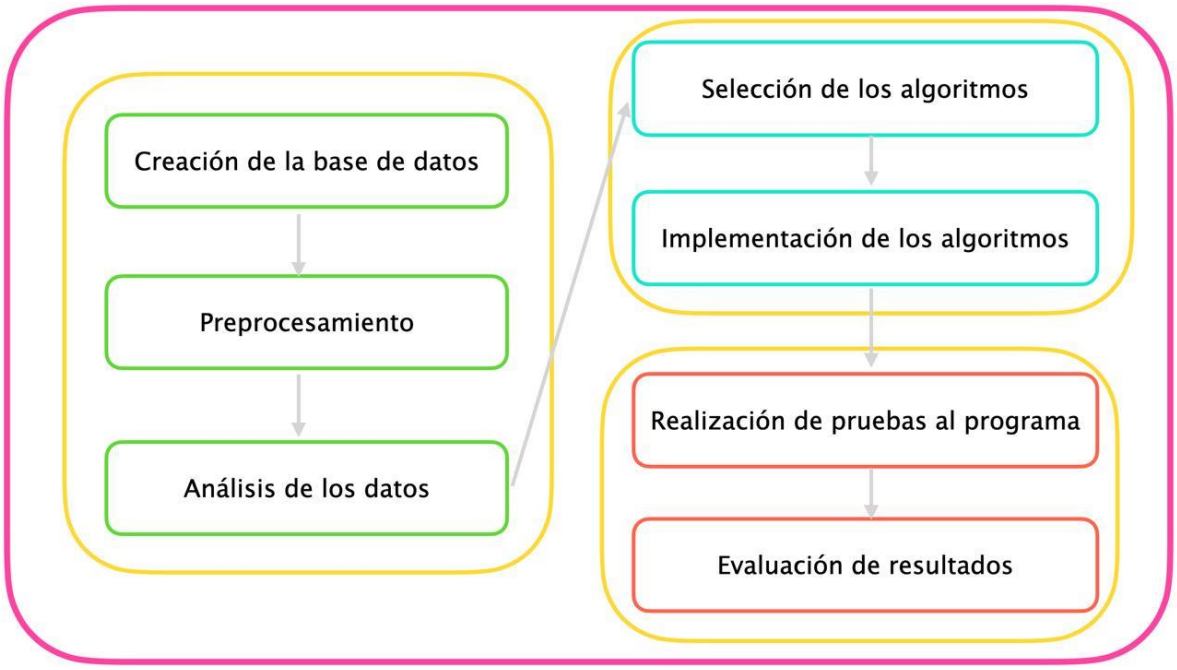

Figura 21. Pasos metodológicos.

Fuente: Elaboración propia.

3.3.1 Creación y análisis de la Base de Datos.

Esta fase responde al objetivo específico de construir una base de datos de exoplanetas preprocesada mediante la identificación y selección de variables clave del sistema planetario. Como salida de esta fase se provee la base de datos limpia y su Análisis Exploratorio de los Datos.

Como primer paso se importan los datos del Archivo de Exoplanetas de la NASA y de la Enciclopedia Exoplanetaria de la NASA donde se obtiene el tipo. Generando así una tabla con variables del sistema planetario y su tipo de exoplaneta.

Una vez obtenida la Base de Datos del paso anterior, se debe realizar un preprocesamiento donde se manejen datos faltantes, remoción de umbral de error, transformación de variables, codificación, etc. En esta fase se obtendrá una Base de Datos limpia.

Con la Base de Datos Limpia se realiza un análisis para determinar la distribución de las variables, su relación entre ellas y obtener la estructura final de la base de datos.

3.3.2 Selección e Implementación de Algoritmos de Aprendizaje Computacional para la Clasificación de Tipo de Exoplaneta.

Esta fase responde al objetivo específico de implementar y ajustar los algoritmos utilizando la base de datos de exoplanetas creada, para clasificar los tipos de exoplanetas en función de las variables del sistema planetario. Como entrada de esta fase se cuenta con la Base de Datos Preprocesada y como salida se obtiene un programa para la clasificación de tipos de exoplanetas.

Se implementan las Redes Neuronales con Aprendizaje Supervisado para la clasificación de tipo de exoplaneta, debido a sus capacidades de aprendizaje en representaciones complejas y no lineales en los datos, logrando extraer las características relevantes para la clasificación. Además, ofrece una amplia flexibilidad en la arquitectura, permitiendo adaptarse a una variedad de problemas y situaciones, ya que pueden establecerse diferentes tipos de capas, funciones de activación, conexiones y arquitecturas. En cuanto a los recursos, es implementada para realizar el entrenamiento en CPU, donde el tiempo de entrenamiento e inferencia depende de la complejidad del modelo y las capacidades de procesamiento de la CPU. El uso de memoria debe contemplar el almacenamiento de los datos (base de datos preprocesada), el modelo entrenado y los resultados

del entrenamiento. El algoritmo se codificará en el entorno COLAB con un entorno de ejecución de 12.7GB de RAM y 107.7 GB de Disco.

La Red Neuronal se codifica en el lenguaje Python debido a que posee las siguientes características: sintaxis limpia y reducida, es gratuito y libre, multiplataforma, lenguaje interpretado, administración automática de memoria, multiparadigma y con la liberaría sklearn que permite la creación de un clasificador de perceptrón multicapa con la función “MLPClassifier”.

3.3.3 Evaluación desempeño de los algoritmos para la clasificación de tipo de exoplanetas.

Esta fase responde al último objetivo específico: evaluar el desempeño del algoritmo de ciencia de datos para la clasificación de exoplanetas, analizando y comparando los resultados utilizando métricas relevantes.

Utilizando el programa generado en la fase anterior se deben probar diferentes escenarios para comprobar el comportamiento de los algoritmos, recolectando los resultados de estas pruebas. Se hará uso de validación cruzada para evaluar el rendimiento de una manera más robusta; utilizando 10 pliegues.

Se deben analizar los resultados obtenidos en el paso anterior determinando el desempeño de cada algoritmo. Estos resultados permiten el ajuste de los algoritmos para mejorar los resultados finales.

3.4 Ajuste de los modelos

3.4.1 Red Neuronal (Aprendizaje Supervisado)

Para la Red Neuronal con Aprendizaje Supervisado se requirió el ajuste de los parámetros expuestos en la Tabla 9.

Parámetro	Significado
learning_rate_init	Controla el tamaño del paso al actualizar los pesos.
solver	El solucionador para la optimización del peso.
activation	Función de activación de la capa oculta.
hidden_layer_sizes	Número de neuronas en la capa oculta
max_iter	Número máximo de iteraciones

Tabla 9. Parámetros Red Neuronal Aprendizaje Supervisado.

Fuente: Elaboración propia.

Para poder ajustar los parámetros, se hicieron 5 rondas, donde inicialmente se colocan los valores por defecto de cada parámetro y en cada ronda se prueban diferentes valores. Los valores pueden ser una lista de opciones o un rango para valores numéricos. Para evaluar los parámetros y evitar el sobreentrenamiento, se hace uso de la validación cruzada, con 10 pliegues, donde se obtiene la mediana y es comparada con la del resto de los parámetros. La mejor precisión es seleccionada y pasa a la siguiente ronda.

El primer parámetro es “learning_rate_init”, debido a que es un valor numérico, se utiliza el rango (0.001, 0.2001) con saltos de 0.001. Debido a la cantidad de entradas, a continuación, se enlistan las mejores 20 precisiones para este parámetro en la Tabla 10. La Figura 22 muestra todas las precisiones obtenidas para este parámetro, donde se aprecia un comportamiento decreciente a medida que el valor del parámetro aumenta.

Valor del Parámetro	Resultados	Precisión
0.016	[0.91, 0.94, 0.89, 0.91, 0.91, 0.9, 0.92, 0.9, 0.92, 0.92]	0.91049325
0.034	[0.92, 0.93, 0.9, 0.92, 0.92, 0.86, 0.9, 0.91, 0.91, 0.91]	0.91176471
0.009	[0.93, 0.92, 0.89, 0.92, 0.91, 0.91, 0.9, 0.9, 0.93, 0.91]	0.91176471
0.024	[0.93, 0.93, 0.9, 0.91, 0.9, 0.91, 0.91, 0.91, 0.93, 0.93]	0.91316527
0.01	[0.93, 0.94, 0.9, 0.92, 0.91, 0.89, 0.9, 0.92, 0.91, 0.93]	0.91316527
0.035	[0.91, 0.89, 0.89, 0.92, 0.91, 0.87, 0.92, 0.91, 0.92, 0.91]	0.91316527
0.015	[0.92, 0.93, 0.89, 0.92, 0.91, 0.89, 0.9, 0.91, 0.95, 0.92]	0.91456583
0.013	[0.94, 0.94, 0.89, 0.91, 0.9, 0.89, 0.93, 0.9, 0.94, 0.92]	0.91456583
0.008	[0.91, 0.93, 0.88, 0.92, 0.91, 0.89, 0.92, 0.9, 0.92, 0.93]	0.91596639
0.006	[0.92, 0.93, 0.9, 0.91, 0.92, 0.89, 0.91, 0.92, 0.94, 0.93]	0.91736695
0.021	[0.89, 0.93, 0.9, 0.92, 0.92, 0.91, 0.91, 0.92, 0.94, 0.94]	0.91736695
0.022	[0.93, 0.93, 0.89, 0.92, 0.9, 0.92, 0.9, 0.89, 0.94, 0.92]	0.91736695
0.004	[0.93, 0.93, 0.9, 0.91, 0.92, 0.89, 0.92, 0.91, 0.93, 0.93]	0.91876751
0.002	[0.92, 0.94, 0.9, 0.91, 0.92, 0.9, 0.92, 0.9, 0.94, 0.93]	0.91876751
0.012	[0.91, 0.93, 0.89, 0.93, 0.92, 0.92, 0.91, 0.9, 0.92, 0.92]	0.91876751
0.001	[0.92, 0.93, 0.88, 0.92, 0.92, 0.91, 0.91, 0.9, 0.92, 0.94]	0.91876751
0.007	[0.92, 0.93, 0.89, 0.92, 0.9, 0.91, 0.93, 0.9, 0.92, 0.92]	0.92016807
0.02	[0.93, 0.93, 0.92, 0.92, 0.92, 0.91, 0.91, 0.91, 0.93, 0.94]	0.92156863
0.005	[0.93, 0.93, 0.9, 0.92, 0.92, 0.89, 0.92, 0.91, 0.95, 0.93]	0.92296919
0.003	**[0.93, 0.94, 0.89, 0.92, 0.93, 0.89, 0.93, 0.91, 0.94, 0.92]**	**0.92577031**

Tabla 10. Precisiones para los valores del parámetro "learning_rate_init" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

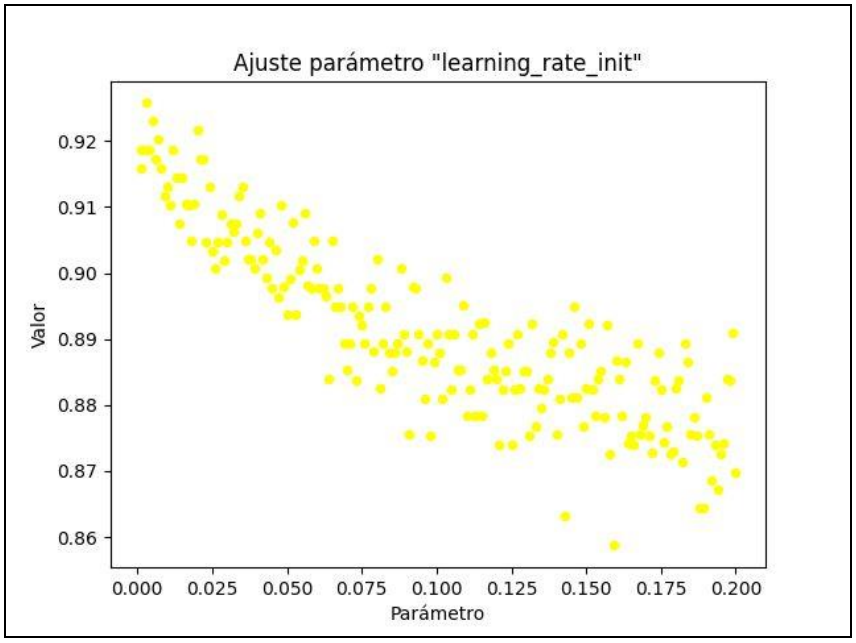

Figura 22. Gráfico de las precisiones para los valores del parámetro "learning_rate_init" en el algoritmo Red Neuronal (Aprendizaje Supervisado)

Fuente: Elaboración propia.

El parámetro "solver" puede tomar los siguientes valores: {'lbfgs', 'sgd', 'adam'}. Las precisiones obtenidas por cada opción para este parámetro se encuentran en la Tabla 11. La Figura 23 muestra el gráfico de estas precisiones; donde el valor que obtiene una mayor precisión es "adam" con una precisión de 0.92577.

Valor del parámetro	Resultados	Precisión
adam	**[0.93, 0.93, 0.9, 0.92, 0.92, 0.9, 0.93, 0.91, 0.94, 0.93]**	**0.92577031**
lbfgs	[0.94, 0.92, 0.89, 0.89, 0.9, 0.9, 0.9, 0.9, 0.91, 0.92]	0.9047619
sgd	[0.9, 0.91, 0.89, 0.89, 0.9, 0.88, 0.91, 0.89, 0.92, 0.89]	0.89495798

Tabla 11. Precisiones para los valores del parámetro "solver" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

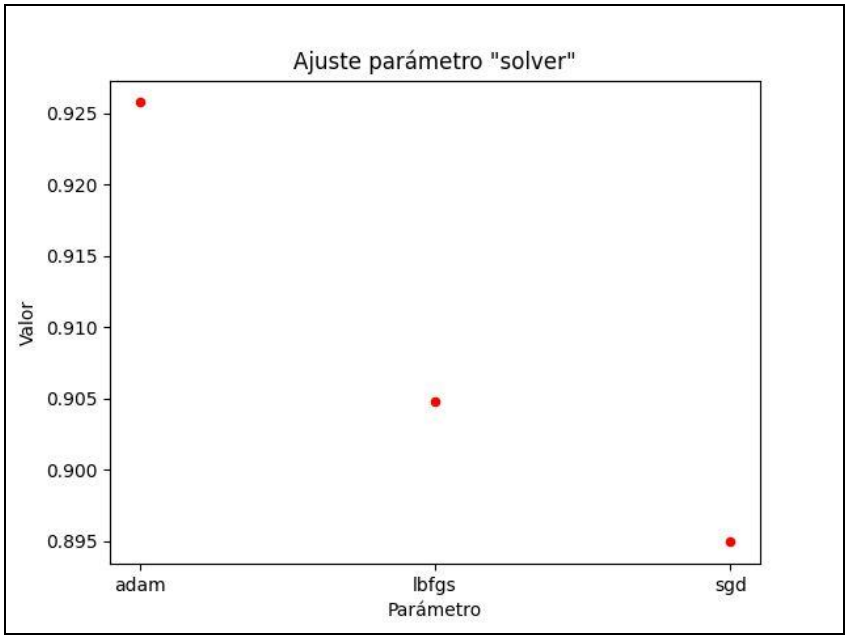

Figura 23. Gráfico de las precisiones para los valores del parámetro "solver" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

Para el caso del parámetro "activation", puede tomar los siguientes valores: {'identity', 'logistic', 'tanh', 'relu'}. Donde las precisiones obtenidas se muestran en la Tabla 12 y su gráfico en la Figura 24. La mejor precisión obtenida para este parámetro es "relu" con 0.923.

Valor del parámetro	Resultados	Precisión
relu	**[0.94, 0.93, 0.89, 0.92, 0.92, 0.9, 0.92, 0.91, 0.94, 0.92]**	**0.92296919**
identity	[0.84, 0.9, 0.85, 0.86, 0.87, 0.85, 0.86, 0.88, 0.92, 0.89]	0.86694678
logistic	[0.92, 0.93, 0.87, 0.92, 0.91, 0.89, 0.9, 0.9, 0.92, 0.92]	0.91607984
tanh	[0.92, 0.95, 0.88, 0.91, 0.91, 0.9, 0.92, 0.92, 0.92, 0.92]	0.91736695

Tabla 12. Precisiones para los valores del parámetro "activation" en el algoritmo Red Neuronal (Aprendizaje Supervisado)

Fuente: Elaboración propia.

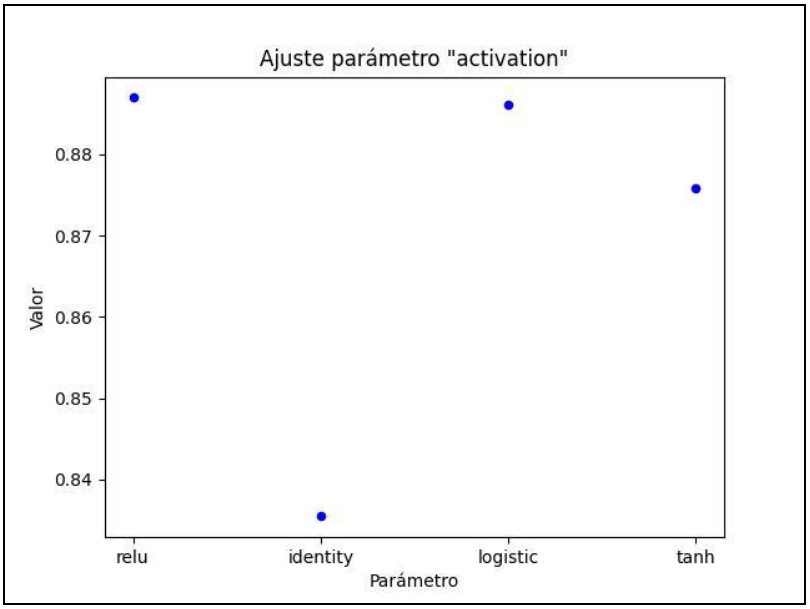

Figura 24. Gráfico de las precisiones para los valores del parámetro "activation" en el algoritmo Red Neuronal (Aprendizaje Supervisado)

Fuente: Elaboración propia.

El cuarto parámetro es "hidden_layer_sizes", un valor numérico que toma los valores dentro del rango (50,3000) con saltos de 50. Las mejores 20 precisiones se encuentran en la Tabla 13 y la Figura 25 muestra el conjunto de valores probado así como las precisiones obtenidas para este parámetro.

Valor del parámetro	Resultados	Precisión
2500	[0.93, 0.93, 0.88, 0.91, 0.92, 0.89, 0.92, 0.91, 0.94, 0.92]	0.92016807
1050	[0.93, 0.94, 0.9, 0.92, 0.93, 0.9, 0.89, 0.91, 0.94, 0.92]	0.92016807
1150	[0.93, 0.92, 0.89, 0.93, 0.94, 0.9, 0.91, 0.92, 0.93, 0.91]	0.9202737
2150	[0.92, 0.92, 0.88, 0.91, 0.92, 0.91, 0.91, 0.92, 0.92, 0.92]	0.92028152
1250	[0.92, 0.94, 0.89, 0.92, 0.92, 0.91, 0.92, 0.92, 0.92, 0.92]	0.92028152
550	[0.94, 0.92, 0.89, 0.91, 0.92, 0.9, 0.91, 0.92, 0.94, 0.92]	0.92156863
900	[0.93, 0.94, 0.9, 0.91, 0.93, 0.9, 0.91, 0.92, 0.93, 0.93]	0.92156863
2250	[0.92, 0.93, 0.89, 0.93, 0.92, 0.9, 0.93, 0.92, 0.93, 0.88]	0.92156863
650	[0.94, 0.92, 0.9, 0.92, 0.92, 0.91, 0.9, 0.92, 0.94, 0.93]	0.92167817
50	[0.92, 0.93, 0.9, 0.92, 0.93, 0.89, 0.92, 0.91, 0.93, 0.93]	0.92168208
300	[0.94, 0.93, 0.89, 0.91, 0.94, 0.9, 0.91, 0.92, 0.92, 0.93]	0.92296919
1700	[0.93, 0.93, 0.89, 0.92, 0.94, 0.91, 0.92, 0.92, 0.92, 0.93]	0.92296919
1550	[0.92, 0.93, 0.89, 0.92, 0.92, 0.91, 0.92, 0.92, 0.93, 0.93]	0.92296919
500	[0.93, 0.92, 0.9, 0.92, 0.94, 0.9, 0.92, 0.91, 0.93, 0.92]	0.92307873
450	[0.93, 0.93, 0.89, 0.92, 0.93, 0.9, 0.92, 0.9, 0.93, 0.93]	0.92436975
1650	[0.9, 0.92, 0.91, 0.92, 0.94, 0.91, 0.94, 0.92, 0.93, 0.92]	0.92436975
1500	[0.93, 0.94, 0.89, 0.92, 0.93, 0.9, 0.91, 0.91, 0.93, 0.93]	0.92436975
750	[0.94, 0.93, 0.89, 0.93, 0.9, 0.89, 0.93, 0.92, 0.92, 0.93]	0.92436975
250	[0.94, 0.94, 0.89, 0.91, 0.93, 0.91, 0.9, 0.92, 0.93, 0.93]	0.92577031
2350	**[0.93, 0.94, 0.89, 0.92, 0.93, 0.93, 0.91, 0.92, 0.93, 0.93]**	**0.92717087**

Tabla 13. Precisiones para los valores del parámetro "hidden_layer_sizes" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

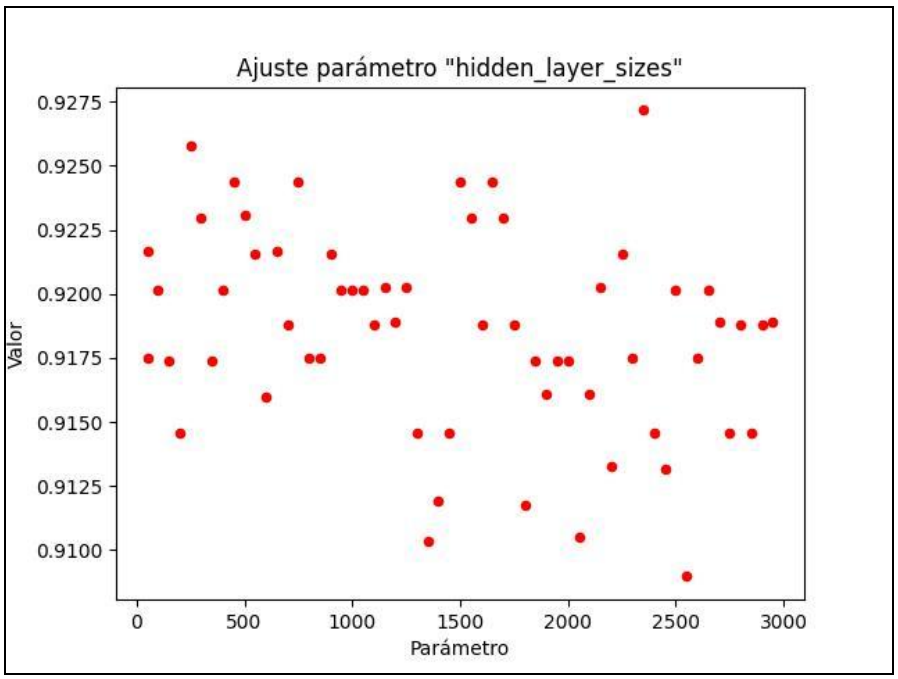

Figura 25. Gráfico de las precisiones para los valores del parámetro "hidden_layer_sizes" en el algoritmo Red Neuronal (Aprendizaje Supervisado)

Fuente: Elaboración propia.

En el caso del último parámetro "max_iter", al igual que "hidden_layer_sizes", se utiliza el rango (50,3000) con saltos de 50; donde las mejores 20 precisiones se encuentran en la Tabla 14. Además, la Figura 26 muestra todas las precisiones obtenidas para este parámetro; en este gráfico no se muestra algún comportamiento general, por lo que la precisión no crece o decrece con algún patrón en particular.

Valor del parámetro	Resultados	Precisión
2700	[0.93, 0.92, 0.89, 0.92, 0.92, 0.9, 0.9, 0.92, 0.92, 0.92]	0.91736695
50	[0.93, 0.94, 0.89, 0.92, 0.92, 0.91, 0.92, 0.91, 0.94, 0.91]	0.91736695
2050	[0.93, 0.92, 0.89, 0.92, 0.92, 0.91, 0.9, 0.92, 0.93, 0.92]	0.91876751
1700	[0.91, 0.94, 0.89, 0.92, 0.89, 0.9, 0.92, 0.92, 0.93, 0.93]	0.91876751
1250	[0.93, 0.93, 0.89, 0.92, 0.91, 0.9, 0.92, 0.92, 0.94, 0.92]	0.91876751
650	[0.92, 0.93, 0.89, 0.91, 0.93, 0.89, 0.92, 0.92, 0.93, 0.93]	0.91888096
2200	[0.93, 0.92, 0.89, 0.9, 0.92, 0.91, 0.92, 0.91, 0.93, 0.93]	0.91888487
1400	[0.92, 0.91, 0.91, 0.92, 0.94, 0.92, 0.92, 0.91, 0.92, 0.93]	0.92016807
2600	[0.94, 0.93, 0.89, 0.92, 0.91, 0.9, 0.92, 0.91, 0.94, 0.93]	0.92016807
2150	[0.94, 0.93, 0.89, 0.92, 0.93, 0.9, 0.92, 0.91, 0.94, 0.91]	0.92016807
950	[0.93, 0.92, 0.89, 0.92, 0.93, 0.89, 0.93, 0.9, 0.92, 0.92]	0.92016807
350	[0.92, 0.93, 0.89, 0.92, 0.94, 0.91, 0.92, 0.92, 0.92, 0.91]	0.92016807
1950	[0.93, 0.95, 0.89, 0.93, 0.89, 0.91, 0.9, 0.91, 0.94, 0.93]	0.92016807
800	[0.93, 0.92, 0.9, 0.92, 0.93, 0.89, 0.91, 0.91, 0.94, 0.93]	0.92027761
2650	[0.92, 0.93, 0.88, 0.92, 0.91, 0.91, 0.93, 0.92, 0.93, 0.92]	0.92028152
1350	[0.94, 0.93, 0.88, 0.92, 0.92, 0.91, 0.92, 0.91, 0.94, 0.92]	0.92156863
1050	[0.93, 0.93, 0.87, 0.92, 0.94, 0.9, 0.91, 0.92, 0.94, 0.92]	0.92156863
1800	[0.92, 0.93, 0.89, 0.9, 0.93, 0.91, 0.92, 0.91, 0.93, 0.92]	0.92156863
2500	[0.94, 0.94, 0.9, 0.9, 0.93, 0.9, 0.9, 0.91, 0.94, 0.94]	0.92296919
1200	**[0.92, 0.93, 0.89, 0.92, 0.93, 0.93, 0.9, 0.92, 0.94, 0.93]**	**0.92447929**

Tabla 14. Precisiones para los valores del parámetro "max_iter" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

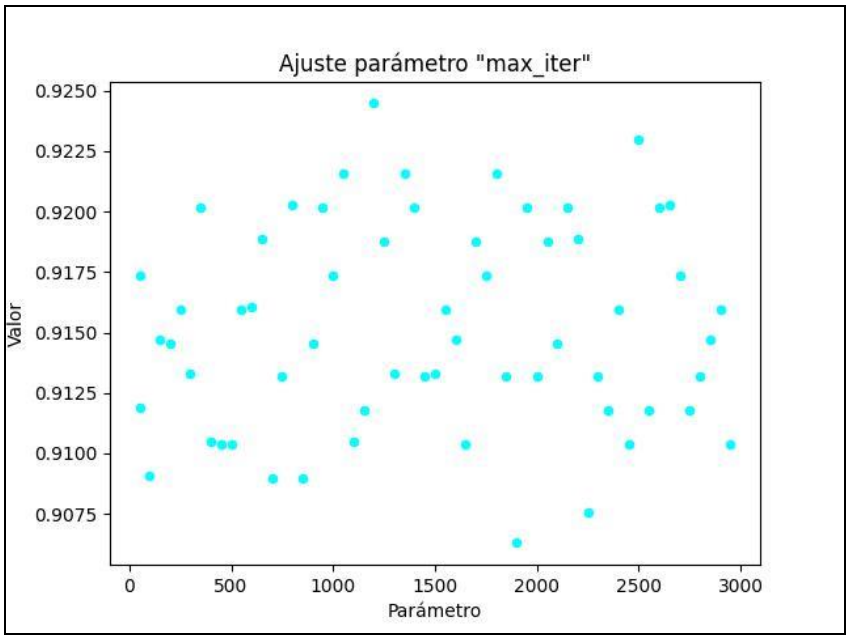

Figura 26. Gráfico de las precisiones para los valores del parámetro "max_iter" en el algoritmo Red Neuronal (Aprendizaje Supervisado).

Fuente: Elaboración propia.

3.5 Análisis de los resultados

3.5.1 Red Neuronal (Aprendizaje Supervisado)

Una vez que se establecen los parámetros, se ejecuta el algoritmo de la Red Neuronal. En la Tabla 15 se muestran las precisiones obtenidas por validación cruzada para el algoritmo de Redes Neuronales.

0.9273743
0.9273743
0.89106145
0.91876751
0.93557423
0.9047619
0.92156863
0.91876751
0.91316527
0.91316527

Tabla 15. Mejores Precisiones del algoritmo Red Neuronal.

Fuente: Elaboración propia.

Utilizando el set de datos de evaluación, se obtiene una precisión de **0.92611**; la matriz de confusión se encuentra en la Tabla 16 y las estadísticas por cada clase se muestran en la Tabla 17. Para el caso de la sensibilidad, se determina que para el algoritmo de Redes Neuronales es más sensible a los positivos en la clasificación del Tipo Terrestre (representado por el 3) y el caso del Tipo Neptuno (representado por el 1) muestra la menor sensibilidad. La Tasa Negativa Verdadera muestra que el tipo Terrestre y Gigante Gaseoso tienen una mayor probabilidad de que un resultado negativo real de un resultado negativo.

El Valor Predictivo Positivo determina que el tipo Terrestre es el que muestra una mayor probabilidad de que al dar un resultado positivo realmente pertenezca a este tipo y los tipos Neptuno y Súper Tierra de exoplaneta muestran una menor probabilidad de acertar. Por otro lado, para el Valor Predictivo Negativo muestra que el tipo Gigante Gaseoso y Terrestre tienen una mayor probabilidad de que al catalogar este no pertenezca a la clase; es decir, nos habla de la seguridad al catalogar; y el tipo Neptuno muestra la menor probabilidad.

	0	*1*	*2*	*3*
0	299	8	1	2
1	11	258	22	2
2	1	30	275	5
3	1	0	5	271

Tabla 16. Matriz de confusión Redes Neuronales.

Fuente: Elaboración propia.

	0	*1*	*2*	*3*
Sensibilidad	0.96	0.87	0.91	0.97
Tasa Negativa Verdadera	0.99	0.96	0.96	0.99
Valor Predictivo Positivo	0.96	0.88	0.88	0.98
Valor Predictivo Negativo	0.99	0.96	0.97	0.99

Tabla 17. Estadísticas por clase Redes Neuronales.

Fuente: Elaboración propia.

Conclusiones y recomendaciones

Conclusiones y recomendaciones

En conclusión, esta investigación se centró en la clasificación de tipos de exoplanetas mediante la implementación de Redes Neuronales utilizando variables del sistema planetario obtenidas de la misión KEPLER. Como resultado del preprocesamiento, se obtuvo una base de datos con 1191 entradas por cada tipo de exoplaneta, es decir, 4764 entradas en total, 9 componentes principales y una variable objetivo.

En el caso de los parámetros se determinó que los siguientes valores mejoran la precisión de las Redes Neuronales para la clasificación de tipo de exoplaneta: 0.003 para el tamaño en la actualización de pesos, “adam” para el parámetro de solucionador, “relu” como activador, 2350 neuronas en la capa oculta y 1200 iteraciones, donde se logró obtener una precisión de 0.92611 al clasificar el tipo de exoplaneta.

En cuanto a las métricas de Sensibilidad, Tasa Negativa Verdadera, Valor Predictivo Positivo y Valor Predictivo Negativo, el Tipo Terrestre obtuvo siempre las probabilidades más altas para estas métricas, seguido por el Tipo Gigante Gaseoso, luego Súper Tierra y finalmente Tipo Neptuno. Este comportamiento puede estar asociado con la aplicación de la herramienta SMOTE, ya que Gigante Gaseoso y el tipo Terrestre fueron las clases con una menor cantidad de entradas, por lo que fueron las más afectadas en la creación de nuevas instancias.

Considero que la gran cantidad de datos que día a día se generan con las misiones espaciales e investigación de exoplanetas, genera una gran variedad de preguntas que sustentan muchos trabajos de investigación. Como una propuesta para la continuación de este trabajo sería la predicción de propiedades de los exoplanetas a partir de su tipo.

Fuentes de consulta

- Agudelo Ovalle, E. P. (2022). Análisis dimensional y caracterización física de exoplanetas. Recuperado de: http://hdl.handle.net/11349/29640.
- Águila, J. A. (2017). Aprendizaje supervisado en conjuntos de datos no balanceados con redes neuronales artificiales: métodos de mejora de rendimiento para modelos de clasificación binaria en diagnóstico médico [Universidad Obera de Cataluya]. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/64768/3/jaguilamaTFM0617memoria.pdf
- Alfaro, E. A. C. (2010). Máquinas de soporte vectorial con algoritmos basados en poblaciones para el pronóstico del precio de acciones LAN Chile. Pontificia Universidad Católica de Valparaıso.
- Arana, C. (2021). Modelos de aprendizaje automoático mediante árboles de decisión (No. 778). Serie Documentos de Trabajo.
- Ball, N. M., & Brunner, R. (2010). DATA MINING AND MACHINE LEARNING IN ASTRONOMY. International Journal of Modern Physics D, 19(07), 1049-1106. https:// doi.org/10.1142/s0218271810017160
- Bustamam, A., Tasman, H., Yuniarti, N., Frisca, & Mursidah, I. (2017). Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). AIP Conference Proceedings. https://doi.org/10.1063/1.4991238
- Capitaine, N., Andrei, A. H., Calabretta, M. R., Déhant, V., Fukushima, T., Guinot, B., Hohenkerk, C. Y., Kaplan, G. H., Klioner, S. A., Kovalevsky, J., Kumkova, I. I., Ma, C., McCarthy, D. D., Seidelmann, P. K., & Wallace, P. T. (2007). DIVISION I / WORKING GROUP NOMENCLATURE FOR FUNDAMENTAL ASTRONOMY. Proceedings of the International Astronomical Union. https://doi.org/10.1017/s1743921308023685
- Carlos, M. B., & De Valladolid Facultad de Ciencias, U. (2022). El método de k-medias. Universidad de Valladolid. https://uvadoc.uva.es/handle/10324/58229
- Castro Becerra, E. F. (2016, septiembre). *Algoritmo de entrenamiento de la red neuronal*. ResearchGate. Recuperado 22 de febrero de 2024, de https://www.researchgate.net/figure/Algoritmo-de-entrenamiento-de-la-red-neuronal-1_fig2_311571158/actions#reference
- Chaviano Arteaga, H. C. (2015). Técnicas de aprendizaje supervisado y no supervisado para el aprendizaje automatizado de computadoras. Memorias del primer Congreso Internacional de Ciencias Pedagógicas, 549-564. https://dialnet.unirioja.es/servlet/articulo?codigo=7192675
- Chinellato, Leandro Micael; Luna, Santiago Hernan; Pera, María Sol; Perren, Gabriel Ignacio; Menchón, Rodrigo Ezequiel; et al.; Clasificación de exoplanetas: desarrollo de una estrategia didáctica para abordar la construcción de modelos observacionales en Física Educativa; Latin American Physics Education Network; Latin American Journal of Physics Education; 16; 4; 12-2022; 1-11

- Curiel, S., & Curiel Ramírez, L. A. (2011, 1 mayo). Búsqueda de exoplanetas. Universidad Nacional Autónoma de México. Dirección General de Cómputo y de Tecnologías de Información y Comunicación. Revista Digital Universitaria. https://www.ru.tic.unam.mx/handle/123456789/1889
- Del Hierro Diez, A. H. D. (2020). ExoplanetIA: Machine Learning para la detecci´on de exoplanetas. UNIVERSIDAD DE VALLADOLID. https://uvadoc.uva.es/bitstream/handle/10324/44403/TFG-G4652.pdf?sequence=1&isAllowed=y
- El Universo, D. (2013). El universo. Obtenido de https://cursa.ihmc.us/rid=1240413396890_1297800289_11749/universo1eso.pdf
- Estupiñan, J. J., Giral, D., & Santa, F. M. (2016). Implementación de algoritmos basados en máquinas de soporte vectorial (SVM) para sistemas eléctricos: revisión de tema. https://www.redalyc.org/journal/2570/257046835012/html/
- Etecé (Ed.). (2021, August 5). Astro - Concepto, tipos y características. Concepto. Retrieved March 21, 2023, from https://concepto.de/astro/
- Furlan, E., Ciardi, D. R., Everett, M. E., Saylors, M., Teske, J., Horch, E., Howell, S. B., Van Belle, G., Hirsch, L. A., Gautier, T. N., Adams, E. R., Barrado, D., Cartier, K., Dressing, C. D., Dupree, A. K., Gilliland, R. L., Lillo-Box, J., Lucas, P. W., & Wang, J. (2017). THEKEPLERFOLLOW-UP OBSERVATION PROGRAM. i. A CATALOG OF COMPANIONS TOKEPLERSTARS FROM HIGH-RESOLUTION IMAGING. The Astronomical Journal, 153(2), 71. https://doi.org/10.3847/1538-3881/153/2/71
- Galindo, E. A., Perdomo, J. A., & Figueroa–García, J. C. (2020). Estudio comparativo entre máquinas de soporte vectorial multiclase, redes neuronales artificiales y sistema de inferencia neuro-difuso auto organizado para problemas de clasificación. InformacióN TecnolóGica, 31(1), 273-286. https://doi.org/10.4067/s0718-07642020000100273
- García Herrero, J., Berlanga de Jesús, A., Patricio Guisado, M. Á., & Padilla, W. (2018). Ciencia de datos: Técnicas analíticas y aprendizaje estadístico en un enfoque práctico.
- García, J. S. (2018). Planetas Extrasolares. Anuario Astronómico del Observatorio, 391-407.
- Garre, M., Cuadrado, JJ, Sicilia, MA, Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. REICIS. Revista Española de Innovación, Calidad e Ingeniería del Software , 3 (1), 6-22.
- Gas Giant | Planet Types – Exoplanet Exploration: Planets beyond our Solar System. (2022, 22 abril). Exoplanet Exploration: Planets Beyond our Solar System. Recuperado 28 de enero de 2024, de https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/gas-giant/
- Giles, D., & Walkowicz, L. M. (2018). Systematic Serendipity: a test of unsupervised machine learning as a method for anomaly detection. Monthly

Notices of the Royal Astronomical Society, 484(1), 834-849. https://doi.org/10.1093/mnras/sty3461

- Goldenberg, D. (2007). Categorización automática de documentos con mapas auto-organizados de Kohonen. https://ri.itba.edu.ar/entities/tesis%20de%20maestr%C3%ADa/c5a9e70a-b9d5-416d-86fd-e55d0e037427
- González Cangrejo, J. G. C. (2021). Algoritmos de Aprendizaje Supervisado en la Clasificación de Exoplanetas en Python. UAN. Retrieved March 8, 2023, from http://repositorio.uan.edu.co/bitstream/123456789/5839/1/2021JohansGonzálezCangrejo.pdf
- González Cangrejo, J. G. C. (2021). Algoritmos de Aprendizaje Supervisado en la Clasificación de Exoplanetas en Python. Universidad Antonio Nariño. http://repositorio.uan.edu.co/handle/123456789/5839
- Hernández, J. A. C. (2003). Exoplanetas: la promesa de una planetología comparada. Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=2898294
- Herrero, J. G., Berlanga, A., López, J., Guisado, M. Á. P., Bustamante, Á. L., & Arias, W. R. P. (2018). Ciencia de datos: técnicas analíticas y aprendizaje estadístico. Alfaomega. https://dialnet.unirioja.es/servlet/libro?codigo=763464
- Lara Torra, J. A. (2011). Modelo para la comparación de datos posturográficos estructuralmente complejos. *UNIVERSIDAD POLITÉCNICA DE MADRID*, *7bf40daf-9453-43f5-86fd-ea0a9f1d6ee6*, file:///Users/monsecampos/Downloads/Modelo_para_la_comparacion_de_datos_posturografico.pdf. http://oa.upm.es/5973/
- Martínez, R. E. B., Ramírez, N. C., Mesa, H., Suárez, I. R., Del Carmen Gogeascoechea-Trejo, M., Pavón-León, P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista Médica de la Universidad Veracruzana, 9(2), 19-24. https://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf
- Maté Jiménez, M. J. (2014). Big Data. un nuevo paradigma de análisis de datos. https://repositorio.comillas.edu/xmlui/handle/11531/4873
- Meléndez Lorenzo, A. M. L. (2022, 20 junio). Estudio, desarrollo y evaluación de técnicas de aprendizaje automático en tareas de clasificación y/o predicción: detección de exoplanetas. Universidad del País Vasco. https://addi.ehu.es/handle/10810/59494
- Molina, F. (2021, 7 junio). Vista de la nueva era de datos en astronomía. Bits de Ciencia. https://revistasdex.uchile.cl/index.php/bits/article/view/1915/1858
- Monzón, C. V. (2021). Detección, dinámica y habitabilidad de exoplanetas y exosatélites (Doctoral dissertation, Universidade de Santiago de Compostela).
- Neptune-like | Planet Types – Exoplanet Exploration: Planets beyond our Solar System. (2023, 10 febrero). Exoplanet Exploration: Planets Beyond

our Solar System. Recuperado 28 de enero de 2024, de https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/neptune-like/

- Overview | Planet Types – Exoplanet Exploration: Planets beyond our Solar System. (2022, 13 abril). Exoplanet Exploration: Planets Beyond our Solar System. Recuperado 28 de enero de 2024, de https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/overview/
- Peláez Chávez, N. P. C. (2012). Aprendizaje No Supervisado y el Algoritmo Wake-Sleep en Redes Neuronales. Universidad Tecnológica de la Mixteca. http://jupiter.utm.mx/~tesis_dig/11612.pdf
- Prithivraj G, & Kumari, A. (2023). Identification and classification of exoplanets using machine learning techniques. arXiv (Cornell University). https://doi.org/10.48550/ arxiv.2305.09596
- Rebolo, R. (2003, 1 enero). Exoplanetas. Rebolo | Revista Española de Física. https://revistadefisica.es/index.php/ref/article/view/591
- Ribas, I. R. (2019, 8 octubre). Exoplanetas, un cosmos lleno de vida. National Geographic. Recuperado 5 de febrero de 2024, de https://www.nationalgeographic.com.es/ciencia/exoplanetas-cosmos-lleno-vida_14789
- Rincón Valencia, K. J. R. V. (2021, 30 junio). Desarrollo de un prototipo de software en Python con técnicas de machine learning para el análisis de datos astronómicos de exoplanetas recopilados por la NASA. Universidad Antonio Nariño. http://repositorio.uan.edu.co/handle/123456789/4472
- Rodríguez, A. R. M., & Gallardo, J. C. (2020). Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil. Dialnet. https://dialnet.unirioja.es/servlet/articulo?codigo=9095465
- Rojas García, E. A. (2016). Análisis de curvas de luz de exoplanetas utilizando datos de la sonda espacial Kepler. Escuela de Ciencias Físicas y Matemáticas. Recuperado 10 de noviembre de 2023, de https://ecfm.usac.edu.gt/sites/default/files/2016-11/Tesis%20Alex%20Rojas.pdf
- Ruiz, N. R., Llorente, I. L., & Domènech Casal, J. D. C. (2017). Vista de Indagación, Exoplanetas y Competencia Científica. Los estudios de Caso como ABP para las Ciencias. Dialnet. Retrieved March 8, 2023, from https://raco.cat/index.php/ECT/article/view/328894/419491
- Salas, R. (2016). Redes neuronales artificiales. uv-cl. https://www.academia.edu/24633757/Redes_Neuronales_Artificiales
- sklearn.neural_network.MLPClassifier. (2024). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- Super-Earth | Planet Types – Exoplanet Exploration: Planets beyond our Solar System. (2022, 13 abril). Exoplanet Exploration: Planets Beyond our Solar System. Recuperado 28 de enero de 2024, de https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/super-earth/
- Terrestrial | Planet Types – Exoplanet Exploration: Planets beyond our solar System. (2022, 13 abril). Exoplanet Exploration: Planets Beyond our Solar

System. Recuperado 28 de enero de 2024, de https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/terrestrial/

- Trujillo, F. R. M., & Cuevas, Z. O. (2006). Prediccion mediante redes neuronales artificiales de la transferencia de masa en frutas osmoticamente deshidratadas. *Interciencia*, *31*(3), 206-210. https://dialnet.unirioja.es/servlet/articulo?codigo=1992175
- Valls, J. M. (2014, 24 septiembre). Simulador de mapas autoorganizados de Kohonen. Universidad Carlos III de Madrid. https://e-archivo.uc3m.es/handle/10016/26282