Algoritmos de Aprendizaje Supervisado en la Clasificación de Exoplanetas en Python

Johans González Cangrejo

Código 20441814958

Universidad Antonio Nariño

Programa de Ingeniería Electrónica

Facultad de Ingeniería Mecánica, Electrónica y

Biomédica Ibagué, Colombia

2021

Algoritmos de Aprendizaje Supervisado en la Clasificación de Exoplanetas en Python

Johans González Cangrejo

Proyecto de grado presentado como requisito parcial para optar al título de:

INGENIERO ELECTRONICO

Director (a):

Ph. D Sergio Alejandro Orjuela Vargas

Codirector (a):

Ph. D Alex Moreno Briceño

Línea de

Investigación:

Electrónica Digital

Universidad Antonio Nariño

Programa de Ingeniería Electrónica

Facultad de Ingeniería Mecánica, Electrónica y

Biomédica Ibagué, Colombia

2021

Contenido

Pág.

Lista de figuras

Lista de tablas

Dedicatoria

A mi familia

Aquellas personas que me han impulsado a lo largo de la vida por el camino recto, sin importar las dificultades presentadas, contratiempos y caídas.
Con su inmenso amor, dedicación y entrega desinteresada, han forjado en mí el valor para continuar con este desafío intelectual abandonado por tantos años y finalmente logrado con sacrificio.
Doy gracias a Dios por contar con su apoyo.

Agradecimientos

El crecimiento intelectual logrado a través del paso por la Educación Superior, demanda por parte del aprendiz la apropiación o cultivo de ciertas habilidades que le permite llevar a feliz término el desafío propuesto. Estas herramientas son alimentadas y reforzadas con el apoyo de los distintos docentes que con su dedicación, esfuerzo y paciencia orientan la formación por el campus. Razón por la cual agradezco al profesor Julián Ospina por su inmensa labor como coordinador del programa de Ingeniería Electrónica, a la Ingeniera Jennifer Triana puente y responsable de mi predilección por el mundo de la programación, al docente Miguel Ángel Montilla quién transmite sus conocimientos con entusiasmo y pasión generando ambientes amenos y agradables al interior de las aulas, al Ingeniero Ricardo Pino por su tiempo extra clase destinado al refuerzo de las distintas prácticas de laboratorio necesarias en la apropiación de conocimientos, al profesor Sergio Orjuela por su labor como Director y en especial a Dr. Alex Moreno Briceño por su trabajo como Codirector, contribución clave para el éxito de esta investigación.

Resumen

Actualmente se cuenta con una gran cantidad de bases de datos, dadas las múltiples fuentes como: redes sociales, movimientos bancarios, consultas en navegadores web de uso particular, empresarial o académico. Un claro ejemplo lo constituye el estudio de exoplanetas realizado por la NASA, a través de múltiples fuentes como observatorios terrestres y telescopios espaciales (NASA, 2021).

Es importante mencionar que, al momento de dar inicio a este trabajo, la base de datos en mención alberga 4512 planetas confirmados; sin duda alguna, una cifra bastante importante con el potencial suficiente para el estudio en busca de patrones y nuevo conocimiento que conlleva a nuevas observaciones.

Las técnicas y tecnologías de Machine Learning[1], a través de sus modelos predictivos, llevan a cabo la tarea de extraer patrones a partir de los cuales se puede obtener un valioso conocimiento que permite tomar mejores decisiones. En este trabajo se implementan algoritmos de aprendizaje supervisado como son: Árbol de Decisiones (Decisión Tree), K-Vecinos más Cercanos (del inglés K Nearest Neighbours, k-NN) y Máquinas de Vectores de Soporte (del inglés Support Vector, SVM) para clasificar planetas extrasolares (exoplanetas) en las clases de Súper Tierra, Gigante Gaseoso, Tipo Neptuno y Terrestre, con los atributos de masa del planeta, radio del planeta, período orbital y eje semimayor.

Los resultados de los diversos modelos predictivos generados con las técnicas y algoritmos de aprendizaje supervisado de ML, realizan una correcta clasificación de los exoplanetas según los tipos de planetas existentes, a partir de los atributos relevantes definidos.

Palabras clave: Machine Learning, validación, exoplanetas, NASA, observatorios, aprendizaje supervisado, algoritmo.

[1] ML, aprendizaje de máquina o aprendizaje automático

Abstract

Currently there is a large number of databases, given the multiple sources such as: social networks, banking movements, consultations in web browsers for private, business or academic use. A clear example is the study of exoplanets carried out by NASA, through multiple sources such as ground-based observatories and space telescopes (NASA, 2021).

It is important to mention that, at the time of starting this work, the aforementioned database contains 4512 confirmed planets; without a doubt, a quite important figure with enough potential to study in search of patterns and new knowledge that leads to new observations.

Machine Learning techniques and technologies, through their predictive models, carry out the task of extracting patterns from which valuable knowledge can be obtained that allows better decisions to be made. In this work supervised learning algorithms are implemented such as: Decision Tree, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) to classify Extrasolar planets (exoplanets) in the Super Earth, Gas Giant, Neptune Type, and Terrestrial classes, with the attributes of planet mass, planet radius, orbital period, and semi-major axis.

The results of the various predictive models generated with ML's supervised learning techniques and algorithms perform a correct classification of exoplanets according to the existing types of planets, based on the relevant attributes defined.

Keywords: Machine Learning, validation, exoplanets, NASA, observatories, supervised learning, algorithm.

Introducción

En el presente, la implementación de grandes bases de datos (Big Data) para el estudio y análisis dentro de diversas áreas del conocimiento, ha estimulado en gran medida el desarrollo de la ciencia de datos, el Big Data y la ciencia de datos se han convertido en un sector estratégico para el análisis de los datos generados en la astronomía por diversos observatorios y ha sido de gran importancia para abarcar correctamente problemas de gran complejidad a la hora de analizar grandes volúmenes de información y obtener todo el conocimiento posible de los mismos. (UNED, 2019).

Este documento se centra en estudiar modelos predictivos con los atributos masa del planeta, radio del planeta, período orbital y eje semimayor para el análisis y clasificación de datos de exoplanetas en los diferentes tipos de planetas existentes que son: Gigante Gaseoso, Tipo Neptuno, Súper Tierra y Terrestre. Estos datos son obtenidos por la NASA, producto de investigación en observatorios terrestres y telescopios espaciales, implementando diversas técnicas y algoritmos de aprendizaje supervisado de ML para generar dichos modelados.

El aprendizaje supervisado de ML, constituye un conjunto de técnicas orientadas a encontrar patrones ocultos en la información y de esta manera encontrar características que diferencien los grupos para realizar predicciones eficientes que permitan clasificar correctamente la información nueva ingresada al algoritmo. (SINNETIC Powering Business Through Science, 2009)

Para poder implementar las diversas técnicas y algoritmos de ML de aprendizaje supervisado, inicialmente es necesario realizar la recolección de la información que se estudiará, que debe ser cualitativa y cuantitativa pues de esto depende el rendimiento del modelo generado.

Además, conocer los atributos y la naturaleza de cada uno de los datos contenidos en el banco de información es de suma importancia a la hora de definir, cuáles serán los atributos relevantes que se ingresarán a los modelados para realizar las pertinentes predicciones requeridas para la clasificación de la nueva información ingresada. (Aprende Machine Learning , 2017).

A continuación, se realiza el pre-procesamiento de la información, siendo uno de los pasos más importantes para generar un buen modelo predictivo, debido a que la calidad de los datos con que se alimenta el modelo determina el rendimiento del algoritmo. Por este motivo es necesario pre-procesarlos correctamente, para tratar situaciones frecuentes en las bases de datos como la perdida de datos causado por medidas no aplicables, espacio en blanco de encuestas e incluso errores humanos que se puedan hallar contenidos en la misma. (Roman, 2019)

Ahora se debe elegir el modelo o el algoritmo de ML a implementar de acuerdo al objetivo que se tenga, en este caso se implementaran técnicas de aprendizaje supervisado como: Árbol de decisión, k-NN y SVM.

Posteriormente, se divide la base de datos en dos partes: conjuntos de entrenamiento, y prueba. Se requiere realizar un entrenamiento del modelo predictivo con los datos de entrenamiento definidos para finalmente probar su eficiencia con los datos de prueba. La idea es que el modelo pueda generalizar correctamente los datos desconocidos y realizar predicciones con exactitud, basado en sus parámetros internos, ajustados mientras el modelo fue entrenado y validado. (Roman, 2019)

Por último, se implementan los algoritmos de ML mencionados, y se realiza un análisis del error, midiendo el rendimiento del modelo generado a la hora de realizar las diversas predicciones y generalizaciones a través de métricas de evaluación, como la eficiencia y la exactitud. Cuando se habla de generalización nos referimos a la capacidad que tienen los modelos de ML de producir buenos resultados o etiquetar correctamente la información al usar datos nuevos. (IArtificial.net, 2021)

Si se desea obtener mejores resultados de los conseguidos inicialmente, se requiere realizar una iteración sobre los pasos anteriores. Con cada iteración, se espera que mejore el entendimiento del problema abarcado y de los datos o la base de datos implementada. Esto permitirá identificar mejores parámetros relevantes para ingresar a los modelos y de esta manera se logra reducir notoriamente el error de generalización que se esté presentando. (IArtificial.net, 2021).

En el capítulo 1 se definió lo que son los exoplanetas, lo que es ML, técnicas de aprendizaje supervisado y métricas de evaluación, además de hablar sobre el proyecto KELT, y diversos satélites y telescopios relevantes de los cuales se extrajo gran parte de la información contenida en la base de datos sobre exoplanetas. En el capítulo 2 se presenta la recolección y filtrado de datos, además del pre-procesamiento de los mismos. En el capítulo 3 se analizan los diversos algoritmos de aprendizaje supervisado de ML. Por último, en el capítulo 4 están las conclusiones y recomendaciones referentes al proyecto desarrollado a lo largo de este documento.

Objetivos

Objetivo general

- Estudiar modelos predictivos con los atributos masa, radio y periodo orbital para la clasificación de exoplanetas a través de técnicas de aprendizaje supervisado de ML.

Objetivos específicos

- Implementar el algoritmo Árbol de decisiones para el análisis de datos de exoplanetas dispuesto por la NASA, producto de investigación en observatorios terrestres y telescopios espaciales.
- Clasificar a través del algoritmo KNN el data set de exoplanetas extraído de la base de datos de la NASA, resultado de la conquista espacial de sondas y telescopios terrestres.
- Implementar modelo SVM para la clasificación de planetas extrasolares descubiertos por telescopios espaciales, observatorios terrestres y recopilados en la base de datos de la NASA.
- Realizar un estudio de cada uno de los algoritmos implementados a través de las distintas métricas de evaluación para determinar qué modelo presenta el mejor rendimiento.

Definiciones

- **Atributos:** Con sinónimos como: características, factor, propiedad o campo. Es la descripción de cada una de las instancias del Data set (González, 2021)
- **Big Data:** Son conjuntos de datos de mayor tamaño y más complejos, procedentes particularmente de nuevas fuentes de datos. (ORACLE, 2021)
- **CNES:** Centro Nacional de Estudios Espaciales. Fundado en 1961, es la agencia gubernamental responsable de dar forma e implementar la política espacial de Francia en Europa. (CNES, 2021).
- **CSV:** (Valores separados por comas). Es un tipo especial de archivo que puede ser creado o editado en Excel. Donde los datos no quedan separados en columnas sino a través de comas, lo que permite ser usados por otros programas. (Microsoft, 2021)
- **DataFrame:** Estructura de datos etiquetada bidimensionalmente, a modo de una hoja de cálculo o tabla SQL. Con frecuencia objeto usado en pandas. (Análisis y visualización de datos usando Python, 2020).
- **Data set:** El conjunto de datos es la materia prima implementada en el desarrollo de un modelo predictivo. (González, 2021)
- **Estrella enana:** Estrella en la fase principal de su evolución, que corresponde desde el nacimiento al agotamiento del hidrogeno en su núcleo. (SEA, 2021)
- **Etiqueta:** En algunas fuentes denominado clase u objetivo, es el factor que queremos predecir a través del modelo predictivo. (González, 2021)
- **Instancia:** También llamado ejemplo o registro, es cada uno de los datos de que se dispone para realizar un estudio en ML. (González, 2021).
- **Inteligencia artificial:** Es la disciplina que trata de crear sistemas capaces de aprender y razonar como un ser humano. (Aura Quantic, 2021).
- **Matplotlib:** Librería de Python especializada en la creación de gráficos en dos dimensiones. (Aprende con Alf, 2020)
- **Modelo predictivo:** Es un conjunto de procesos llevados a cabo en equipos de

cómputo, para inferir la probabilidad de que ocurran situaciones previas a su ejecución. (B12, 2020)

- **Pandas:** Librería de Python para trabajar con datos tabulares, que provee estructuras para estos Data Sets. (Pandas, 2021)
- **Procesamiento de Lenguaje Natural (PLN):** Es el campo de conocimiento de la Inteligencia Artificial que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino. (Instituto de ingenieria del conocimiento, 2021)
- **Seaborn:** Biblioteca de Python para la visualización de datos con interfaz de alto nivel para gráficos estadísticos. (Seaborn, 2021)
- **Training Data:** Los datos de entrenamiento, son los datos usados para entrenar un modelo predictivo. De su calidad depende el rendimiento del algoritmo. (Santos, 2020)
- **Test Data:** Datos de prueba son los datos que existen antes de que una prueba sea ejecutada y que afectan o son afectados por el componente o sistema en pruebas. (Globe, 2018)
- **UAI:** Unión Astronómica Internacional. Órgano de decisión internacional en la definición de cuerpos celestes y los estándares de Astronomía. (Astropedia, 2021)
- **WFCAM:** Cámara de campo amplio de UKIRT. (Universidad de Arizona, 2021)

Capítulo 1

1.1 Marco teórico

1.2 Exoplanetas

Son denominados planetas extrasolares o exoplanetas, aquellos planetas que orbitan una estrella ajena a la nuestra (Sol). El estudio de estos cuerpos celestes tiene connotaciones filosóficas, pues se busca dar respuesta a las preguntas que infinidad de seres a lo largo del devenir humano se ha planteado. ¿Existe vida inteligente más allá del sistema solar? ¿Acaso hay más planetas habitables? (Ruiz, 2017)

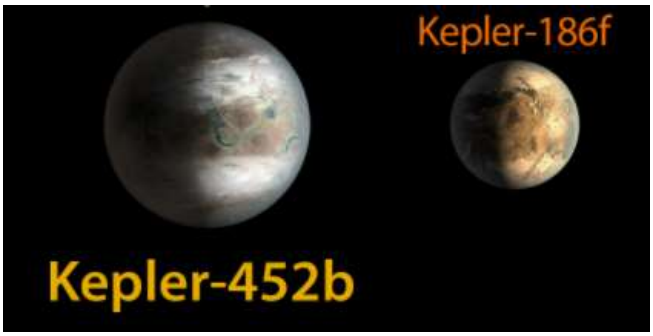
A pesar de las inmensas distancias que los separada de este mundo, es posible determinar algunas características geológicas y climáticas gracias a los métodos de detección, implementados es su búsqueda (tránsito, bamboleo estelar, velocidad radial), denominadas técnicas indirectas (Ruiz, 2017).

Los exoplanetas son denominados a través de una designación científica con los siguientes elementos:

- Un nombre propio (o abreviatura) acompañado con frecuencia de números.
- Una letra minúscula.

La fuente del nombre propio puede ser el nombre de la estrella anfitriona o el instrumento de descubrimiento. Como ejemplo se tiene el exoplaneta 51 Pegasi b, que orbita la estrella 51 Pegasi, o los planetas extrasolares Kepler-69c o CoRoT- 7b descubiertos por el telescopio de la NASA y la sonda espacial Francesa del Centro Nacional de Estudios Espaciales. Ver figuras 1-1 y 1-2 (UAI, 2019).

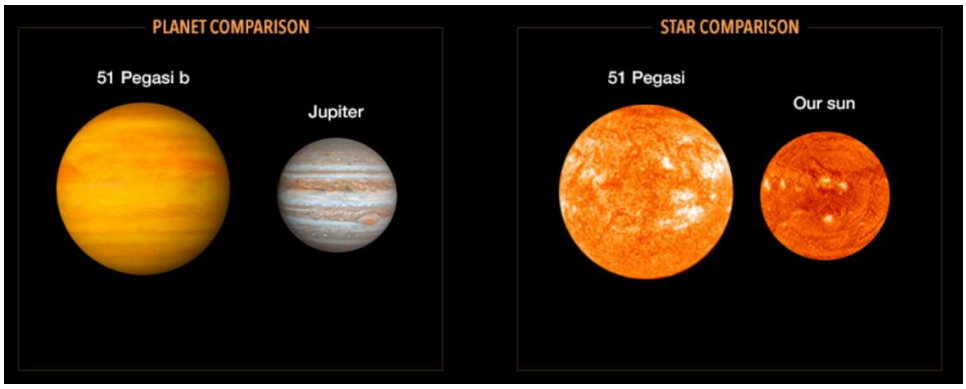
Figura 1-1: Designación científica por instrumento.



Nombre de la fuente: NASA **Recuperado de:**
https://www.nasa.gov/sites/default/files/thumbnails/image/452b_system_comparison.jpg

La nomenclatura de letras minúsculas para el caso del segundo elemento, está definida por la Unión Astronómica Internacional, esta letra es aplicada a todos los estilos de designación e indica el orden de descubrimiento del exoplaneta sobre su estrella anfitriona.

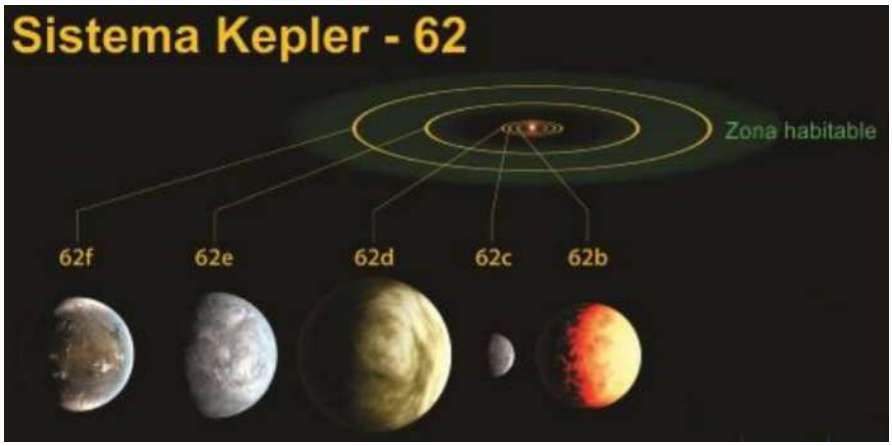
Figura 1-2: Designación científica por estrella anfitriona.



Nombre de la fuente: NASA **Recuperado de:**
https://exoplanets.nasa.gov/resources/289/infographic-profile-of-planet-51-pegasi-b/

El primer planeta extrasolar es designado con la letra minúscula “b”, el segundo “c”, el tercero “d” etc., lo que no indica, su cercanía con respecto a su estrella, razón por la cual un exoplaneta “c” puede estar ubicado antes de uno “b” o después del “d”. Ver figura 1-3 (UAI, 2019).

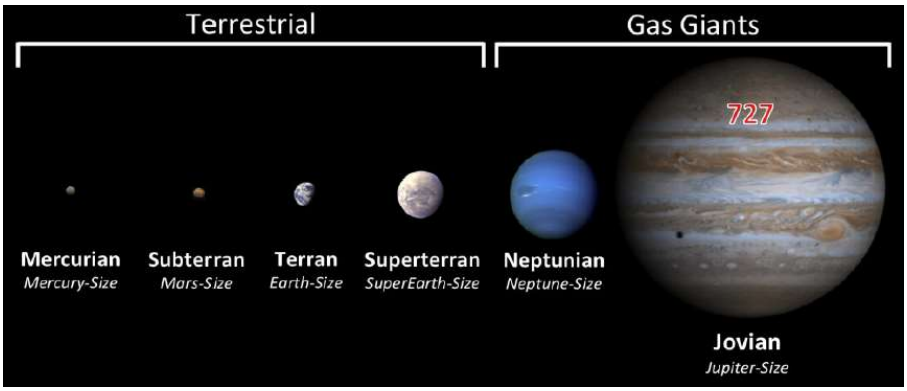
Figura 1-3: Designación de lera minúscula.



Nombre de la fuente: Universidad Politécnica de Cataluña. **Recuperado de:** http://sac.csic.es/astrosecundaria/es/cursos/formato/materiales/conferencias_talleres/T9_es.pdf

Gracias al estudio realizado por los investigadores, los exoplanetas pueden ser clasificados en las siguientes clases: Gigante Gaseoso (con un tamaño superior al de Júpiter y Saturno), Tipo Neptuno, Súper Tierra y por último los Terrestres (planetas sólidos con atmosferas similares a la de Venus o la Tierra y posiblemente con agua superficial). Ver figura 1-4. (García, 2018).

Figura 1-4: Tipos de exoplanetas.



Nombre de la fuente: Anuario Astronómico del Observatorio de Madrid. **Recuperado de:** http://astronomia.ign.es/rknowsystheme/images/webAstro/paginas/documentos/Anuario/articuloSolis2018-apares.pdf

A continuación, se relacionan los telescopios espaciales, naves y observatorios terrestres relevantes en el descubrimiento de exoplanetas.

Kilodegree Extremely Little - KELT

El proyecto KELT consta de dos telescopios instalados en el hemisferio Norte y Sur África, el cual realiza un estudio fotométrico para exoplanetas en tránsito, donde se toman imágenes de decenas de miles de estrellas durante la noche, con el objeto de capturar los exoplanetas al momento de pasar frente a la estrella que orbitan.

Su objetivo se ha centrado en los "Júpiter calientes", cuerpos tan grandes como el Júpiter del Sistema Solar, pero con un periodo orbital entre 2 a 10 días lo que explica sus altas temperaturas dada su cercanía a la estrella que orbitan, al darse el tránsito la intensidad de la estrella es atenuada en un 1%, tarea en la cual se especializa esta investigación. Ver figuras 1-5 y 1-6 (Kilodegree Litle, 2019).

Figura 1-5: Telescopio instalado en el hemisferio Norte.

Nombre de la fuente: Enciclopedia Universo. **Recuperado de:** https://enciclopediauniverso.com/el-universo-es-enorme/kelt-9b-el-exoplaneta-mas-caliente-de-jupiter-hasta-ahora

Figura 1-6: Telescopio instalado en Sur África.

Nombre de la fuente: LEHIGH University. **Recuperado de:** https://www2.lehigh.edu/news/are-we-alone-in-the-universe

Transiting Exoplantet Survey Satellite - TESS

El satélite de estudio de exoplanetas en tránsito, fue concebido para el descubrimiento de planetas extrasolares, orbitando las estrellas enanas más brillantes del firmamento, así mismo, por dos años realiza su misión principal, monitoreando el brillo de estas estrellas en busca de caídas periódicas generadas por tránsitos planetarios con la captura de imágenes del 75% del cielo estrellado, actualmente rastrea pequeños exoplanetas rocosos y planetas gigantes como misión extendida.

Al momento de escribir estas líneas en el registro de exoplanetas de la NASA, se contaba con 74 planetas confirmados y 1200 candidatos. Ver figura 1-7 (NASA, 2021)

Figura 1-7: Satélite para exoplanetas en tránsito.

Nombre de la fuente: NASA. **Recuperado de:** https://www.nasa.gov/content/goddard/nasa-s-tess-mission-cleared-for-next-development-phase

United Kingdom InfraRed Telescope - UKIRT.

El telescopio infrarrojo del Reino Unido, se encuentra catalogado como uno de los más grandes del mundo en su tipo, con rango de longitud de onda entre 1 a 30 micrómetros, conforma los observatorios de Maunakea (Isla Grande de Hawai), así mismo, es financiado por la NASA, operado conjuntamente por el Centro Tecnológico Avanzado Lockheed Martin, la Universidad de Hawai y el Observatorio Naval de EEUU. Ver figura 1-8.

Cuenta con la cámara de campo amplio de infrarrojos (WFCAM), considerado como el mejor instrumento de estudio de imágenes de infrarrojos del hemisferio norte. (Universidad de Cambridge, 2021)

Figura 1-8: Panorama del telescopio infrarrojo.

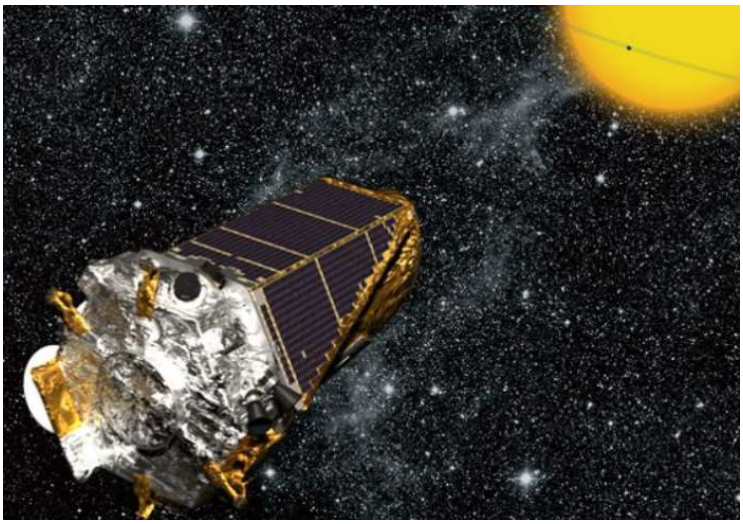
Nombre de la fuente: UKIRT. **Recuperado de:**

https://about.ifa.hawaii.edu/ukirt/about-us/

Telescopio espacial Kepler

Despega el 6 de marzo de 2009 a las 7:49 PM, hora de verano del pacifico, a bordo del cohete Delta II, en la Estación de la Fuerza Aérea de Cabo Cañaveral (Florida). Se conoce como "La primera nave espacial de caza de planetas". Su trabajo consiste en la búsqueda de exoplanetas similares a la Tierra, en un área espacial con un estimado de 150.000 estrellas como el Sol, lo que le brinda ventajas sobre el telescopio Hubble y observatorios terrestres. Sus detectores especializados cuentan con tecnología de cámaras digitales para aprovechar la técnica de Tránsito de observación. Ver figura 1-9. (Nasa, 2018).

Figura 1-9: Nave espacial Kepler (interpretación artística).



Nombre de la fuente: NASA. **Recuperado de:**

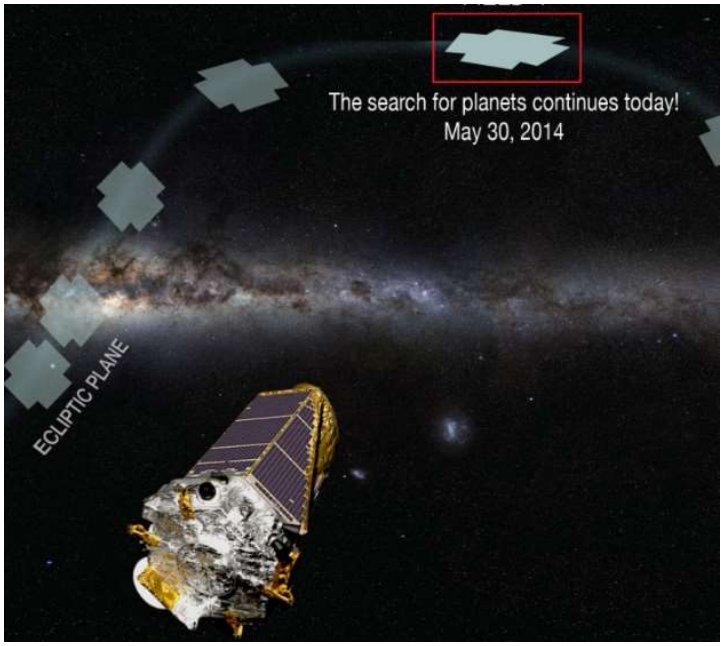
https://spaceplace.nasa.gov/all-about-exoplanets/sp/

Misión K2

Inició en 2015, haciendo uso de la nave espacial Kepler, cuya misión queda suspendida debido a falla en la rueda de reacción, continúa la observación de supernovas, galaxias, estrellas y exoplanetas por períodos de 80 días, gracias a la solución innovadora para orientar el telescopio mediante la gestión de la presión solar y el uso de propulsores.

Se destaca por confirmar 300 exoplanetas y contar con 500 candidatos. Ver figura 1-10 (Seti, 2021) (Ball, 2020).

Figura 1-10: Nave espacial Kepler, operando como misión K2.



Nombre de la fuente: NASA. **Recuperado de:**
https://www.nasa.gov/mission_pages/kepler/overview/index.html

1.2.1 Técnicas de detección

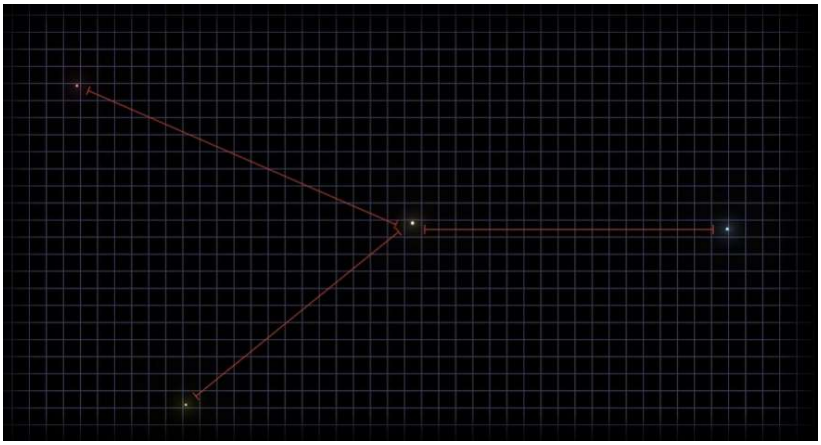
Los cinco métodos implementados en la detección de exoplanetas son:

- Astrometría.
- Microlentes gravitacionales.
- Imagen directa.
- Velocidad radial.
- Tránsito. (NASA, 2021)

Astrometría

Entre los métodos de detección, es considerado el más antiguo, consiste en medir la posición de las estrellas a lo largo del tiempo, actividad realizada por la humanidad aun sin contar con equipo tecnológico apropiado, es de considerar, que un planeta extrasolar genera un movimiento minúsculo en su estrella anfitriona, por lo que el descubrimiento de estos cuerpos a través de este método es potencialmente bajo, no obstante se ha constituido como herramienta complementaria a las demás técnicas de detección. Ver figura 1-11. (REYES, 2014)

Figura 1-11: Método de Astrometría.



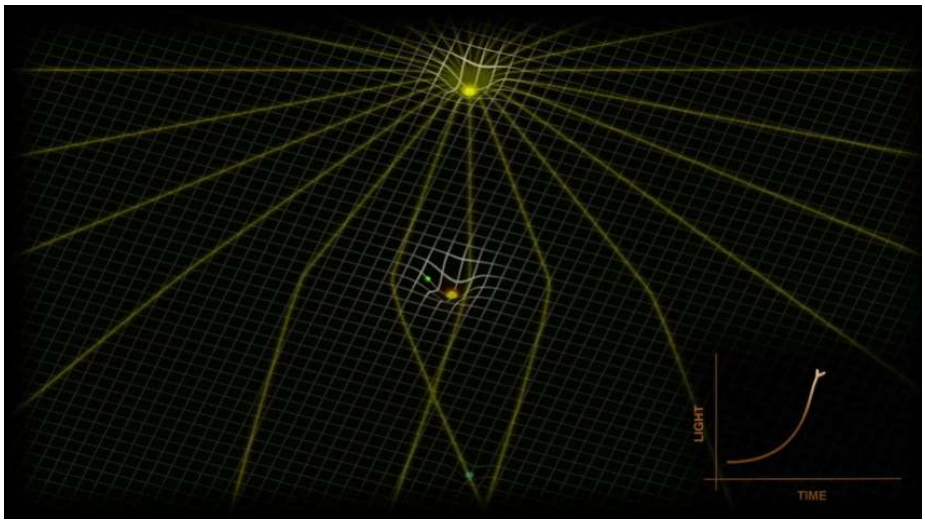
Nombre de la fuente: NASA. **Recuperado de:**
https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/5

Microlentes gravitacionales

Ante la presencia de cuerpos celestes masivos, se genera una curvatura del espacio-tiempo, como lo afirmaba en su famosa teoría Albert Einstein. Con el objetivo de aprovechar este método, es necesario realizar una constante observación de extensas zonas espaciales con una correcta alineación del observatorio y las estrellas, buscando curvas de luz con incrementos en luminosidad, lo que indicaría la presencia de un elevado número de estrellas. Ver figura 1-12.

A través de este método, se registra el primer descubrimiento en 2003 a 19000 años luz y actualmente se han detectado 78 exoplanetas, ahora bien, como dificultad de la técnica se menciona la posibilidad de observar la alineación una sola vez, lo que obliga a implementar otros métodos para confirmar los descubrimientos. (Sánchez, 2019)

Figura 1-12: Luz en una lente de gravedad.



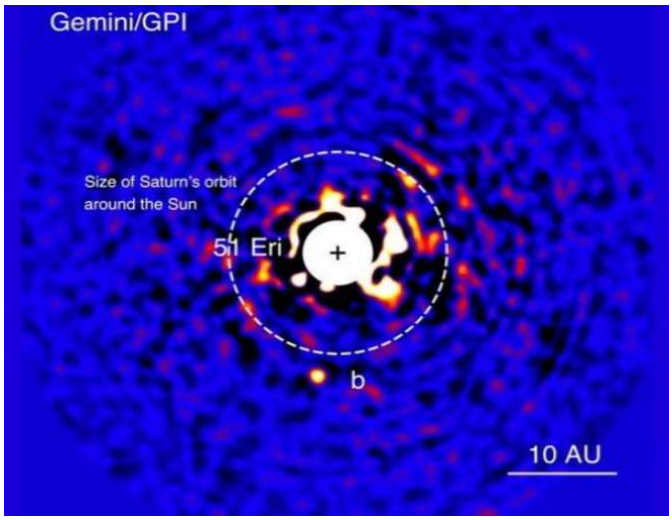
Nombre de la fuente: NASA. **Recuperado de:**

https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/4

Imagen directa

A través de este método se logra la captura directa de estos cuerpos celestes. Como se puede apreciar, la principal dificultad a superar, radica en el brillo excesivo de su estrella anfitriona que "ahoga" la luz procedente del exoplaneta impidiendo su observación, para superar lo anterior se implementan dos técnicas como son: uso de equipo especializado para bloquear la superficie brillante de la estrella y observación de longitudes de onda infrarrojas. Ver figura 1-13. (Esero spain, 2021)

Figura 1-13: Imágenes directas.



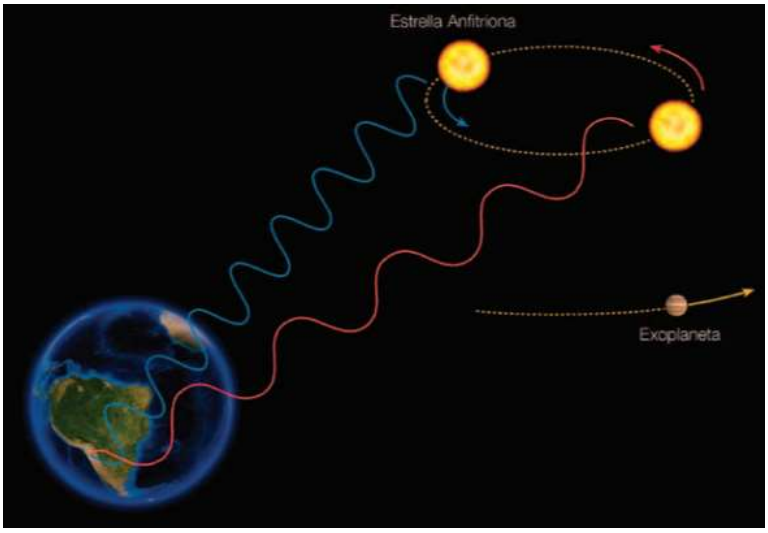
Nombre de la fuente: Universidad Politécnica de Cataluña. **Recuperado de:** http://sac.csic.es/astrosecundaria/es/cursos/formato/materiales/conferencias_talleres/T9_es.pdf

Velocidad radial

Es posible descubrir información relevante de estrellas distantes, a través del estudio de su espectro, esto debido a la fuerza gravitatoria estrella-exoplaneta. La estrella anfitriona se mueve en una pequeña orbita (es un error pensar que las estrellas se encuentran inmóviles en los sistemas planetarios), lo que la aleja o acerca del laboratorio de observación (Tierra), esta velocidad radial provoca que su espectro de longitud de onda oscile entre tonalidades rojas y azules, gracias al efecto Doppler. Ver figura 1-14.

Con espectrógrafos de alta precisión, los investigadores estudian estos desplazamientos Doppler, buscando variaciones periódicas para determinar la masa del exoplaneta. El HARPS (Buscador de planetas por velocidad radial de alta precisión) instalado en Chile, es catalogado como el instrumento más productivo en este campo. (ESO, 2021)

Figura 1-14: Método velocidad radial.



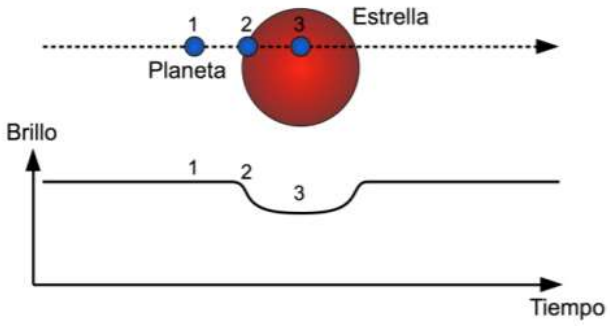
Nombre de la fuente: ESO. **Recuperado de:**
https://www.eso.org/public/archives/presskits/pdf/presskit_0004.pdf

Tránsito

Este método, es por mucho, el más implementado en la exploración de planetas extrasolares, su principio radica, en medir el cambio de intensidad luminosa de una estrella al paso de un cuerpo celeste, un fenómeno como el experimentado en el eclipse solar, es así, como al percibir este aumento y diminución de intensidad de forma periódica, los investigadores pueden concluir que se debe a un objeto orbitando la estrella. Ver figura 1-15.

Información adicional, puede ser determinada por este método como tamaños, masas y distancia estrella-exoplaneta. Al momento del tránsito, la luz estelar atraviesa la atmosfera del exoplaneta permitiendo determinar los elementos que la componen. (Los exoplanetas, 2021)

Figura 1-15: Exoplaneta en tránsito.



Nombre de la fuente: Universidad Politécnica de Cataluña. **Recuperado de:** http://sac.csic.es/astrosecundaria/es/cursos/formato/materiales/conferencias_talleres/T9_es.pdf

1.3 Machine Learning (aprendizaje automático, ML)

ML es una rama derivada de la Inteligencia Artificial, a través de la cual los equipos de cómputo pueden llevar a cabo una tarea sin necesidad de programación explícita, utilizando código escrito, una maquina ejecuta una acción siguiendo reglas específicas, como una secuencia de pasos. Por su parte ML, descubre las reglas detrás de unas salidas esperadas ante unos datos de entrada (Marrugat, 2020).

En 1943 el matemático Walter Pitts y el neurofisiólogo Warren McCulloch sentaron las bases de la Inteligencia Artificial, al llevar a cabo estudios con el objeto de desarrollar equipos que trabajasen al igual que el cerebro humano. Para 1952 Arthur Samuel, escribe un programa de damas chinas capaz de aprender y mejorar su juego partida tras partida. En esta misma década el termino ML fue usado por primera vez, actualmente ha ganado protagonismo gracias al incremento del poder de cómputo y el Big Data (Pacheco, 2019) (Iberdrola, 2021).

Su campo de acción se extiende a ramas de la ciencia como es la Medicina, donde el equipo de cómputo adquiere la capacidad de diagnosticar según las mediciones del tejido entre tumores malignos o benignos para el tratamiento de cáncer de mama (Muller & Guido, 2017). En el campo de la Robótica se han desarrollado trabajos de investigación como es la resolución de cubos de Rubik y medidas antifraude con tarjetas de crédito a nivel bancario. (BBVA, 2019).

El desarrollo de vehículos autónomos es un tema cotidiano, donde se estima que en 2025 se contará con este tipo de servicio en carreteras, con prestaciones adicionales de confort (temperatura, inclinación de respaldo y música a gusto del usuario). Además, se cuenta con asistentes virtuales como Alexa, PLN, capacidad de interpretar sentimientos y traducción entre idiomas. (Iberdrola, 2021).

Para lograr todo lo anterior, es necesario llevar a cabo el estudio o análisis del Big Data proveniente de múltiples fuentes como redes sociales, navegadores web, sensores,

instituciones académicas para investigación, innovación y desarrollo del sector público y privado, estos conjuntos de datos son de naturaleza multivariable, lo que desafía el uso de técnicas estadísticas analíticas o manuales en la deducción de patrones.

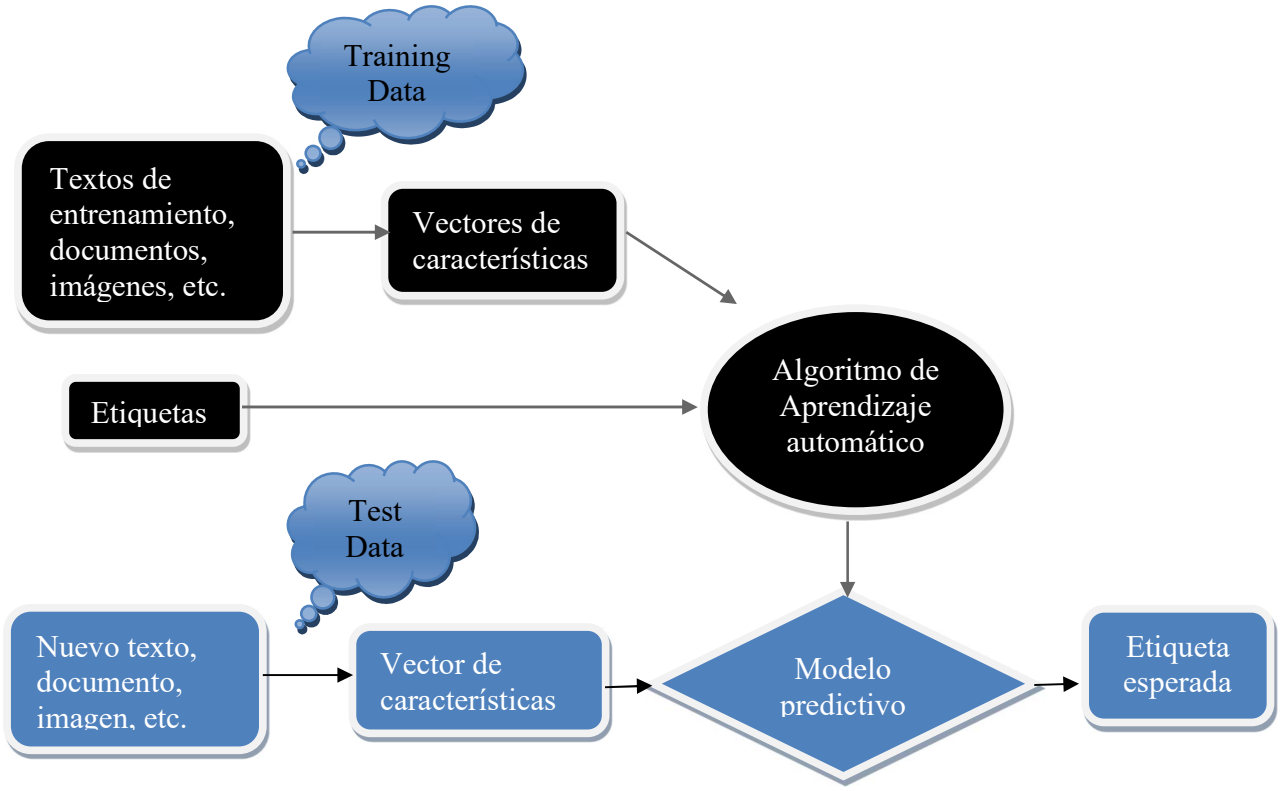
El ML a través de sus potentes técnicas o modelos explota el Big Data, en busca de patrones ocultos extrayendo nuevo conocimiento y mejorando la toma de decisiones ante evento futuros, estos modelos predictivos se clasifican en las categorías de aprendizaje supervisado (centrando el desarrollo del presente trabajo en su implementación), aprendizaje no supervisado y aprendizaje por refuerzo.

1.3.1 Aprendizaje supervisado (Supervised Learning)

El aprendizaje supervisado es la categoría más usada en el ML, donde se cuenta con ejemplos de pares entrada-salida. Por tal motivo, es implementado cuando se requiere hacer predicciones precisas a partir de nuevas entradas no vistas por el modelo (Muller & Guido, 2017). El desarrollo de técnicas a través de esta corriente opera a través del training data para realizar el ajuste del modelo, el cual contiene unos atributos o características bien definidas para realizar la predicción en etiquetas (Dipanjan Sarkar, Bali, & Sharma, 2018).

La ilustración 1-16 presenta la descripción gráfica del aprendizaje supervisado. Como se puede apreciar el modelo es entrenado con el training data y las etiquetas de clasificación. A continuación, se realizan unas predicciones a través del Test Data para proceder como última instancia a probar el algoritmo con el apoyo de las métricas de evaluación, tema a estudiar en próximas líneas.

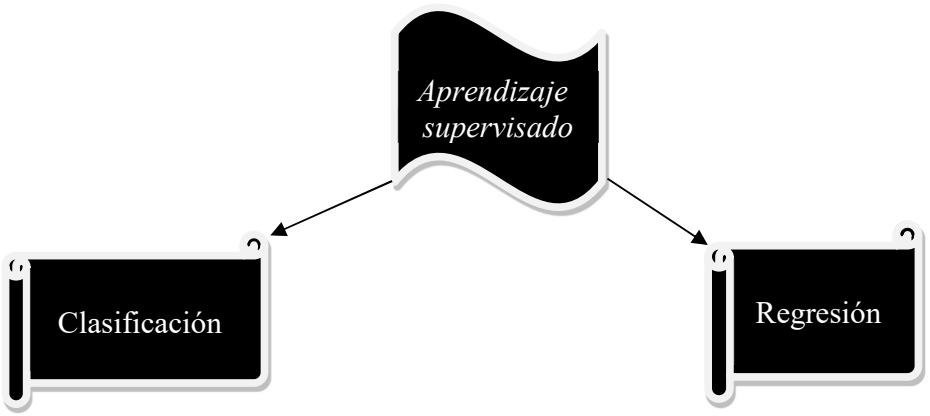
Figura 1-16: Esquema general de aprendizaje supervisado.



Es posible comprender el aprendizaje supervisado intuitivamente, con el siguiente ejemplo: un docente supervisa el proceso de formación a través de unos objetivos definidos al inicio del periodo académico (etiquetas), si el alumno no logra estas metas, el profesor realiza tutorías frecuentes (iteraciones en el desarrollo del modelo) hasta lograr el resultado esperado (predicciones precisas). (Serra, 2020)

Se pueden presentar dos tipos de problemas en aprendizaje supervisado denominados clasificación y regresión como se expone en la figura 1-17.

Figura 1-17: Tipos de problemas en aprendizaje supervisado.



Ante un problema de clasificación, la meta consiste en predecir una etiqueta de clases definidas en el entrenamiento, un ejemplo claro lo constituye la predicción de correos en las etiquetas spam o no spam, según parámetros de entrada definidos. Además, se cuenta con dos tipos de clasificación: binaria y multiclase. Como binaria se tiene la clasificación de correos mencionada, por su parte en el multiclase se cuenta con más de una clase para realizar la predicción.

Ahora bien, en los problemas de regresión se busca predecir un número continuo (punto flotante en programación). Como casos prácticos se tiene la predicción de ingresos anuales por núcleo familiar, según el nivel educativo de sus intrigantes, edad y experiencia laboral o la producción agropecuaria cuyas características pueden ser el estimado de rendimientos históricos, condiciones climáticas y colaboradores, es claro que el valor previsto en estos casos es un número dentro de un intervalo definido. (Muller & Guido, 2017)

Gracias a lo expuesto, es posible concluir que el caso que ocupa el presente trabajo deber ser abordado como un problema de clasificación multiclase al contar con las clases bien definidas de tipos de planeta, a continuación, se exponen las técnicas de aprendizaje supervisado a ser modelados.

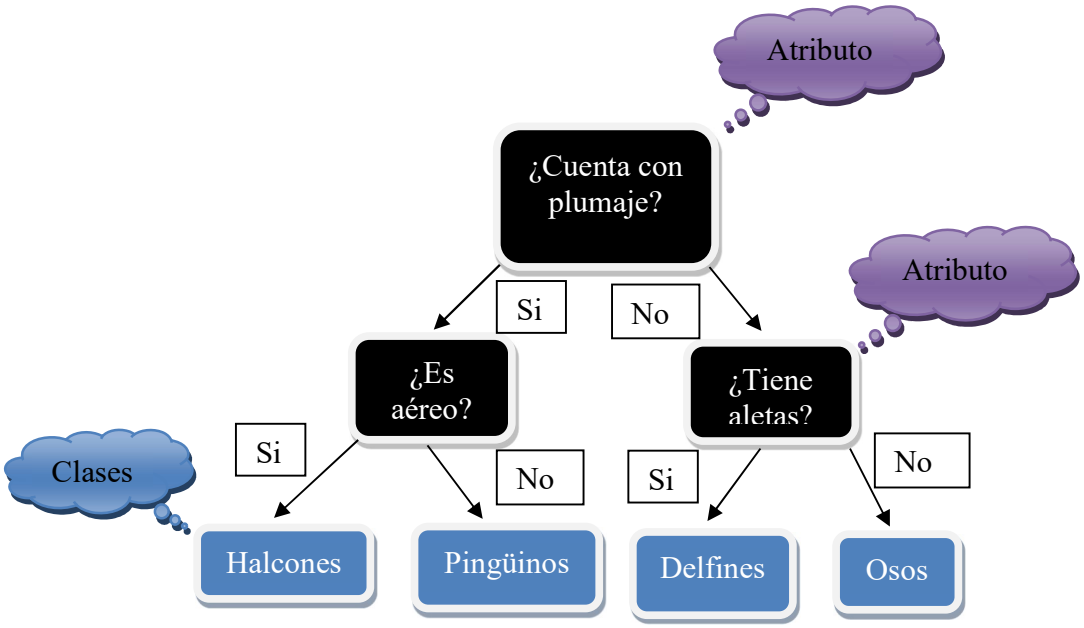
Árboles de Decisiones

El algoritmo basado en árbol de decisiones es considerado de los mejores y más utilizado en aprendizaje supervisado, para tareas de clasificación y regresión, su nombre se debe a su estructura de árbol a modo de diagrama de flujo. Un nodo interno representa una característica también conocida como atributo, las ramas representan una regla de decisión y los nodos-hojas el resultado. De esta manera la clasificación se realiza por el árbol desde la raíz (nodo interno) hasta algún nodo-hoja imitando la operación realizada por el pensamiento humano. (Serra, 2020)

Este modelo se interpreta como una jerarquía de preguntas del tipo if-else, concluyendo en una decisión. Para ilustrar su operación, es posible recrearlo en un juego típico, donde es necesario clasificar cuatro tipos de animales a base de preguntas. Por ejemplo, se cuenta con las etiquetas “osos”, “halcones”, “pingüinos” y “delfines”, al cuestionar, ¿El animal tiene plumas?, se deduce las posibilidades entre halcones y pingüinos en caso contrario osos y delfines. En esa misma línea se podría preguntar si ¿vuela?, donde la respuesta afirmativa seria halcones quedando excluida la etiquita pingüinos. Por otro lado, entre “osos” y “delfines”, sería recomendable averiguar si posee aletas para realizar la clasificación en “delfines”, de esta manera el modelo termina en la cuarta clase con la predicción “osos”. (Muller & Guido, 2017)

Este proceso pregunta-decisión genera una representación tipo árbol, como se aprecia en la figura 1-18. En cada nodo se ejecutan evaluaciones de características, que se asocian a más preguntas. La ilustración pone de relieve las cuatro clases “osos, halcones, pingüinos y delfines” y los tres atributos “plumas, capacidad de vuelo y aletas”.

Figura 1-18: Algoritmo árbol de decisión caso ilustrativo.



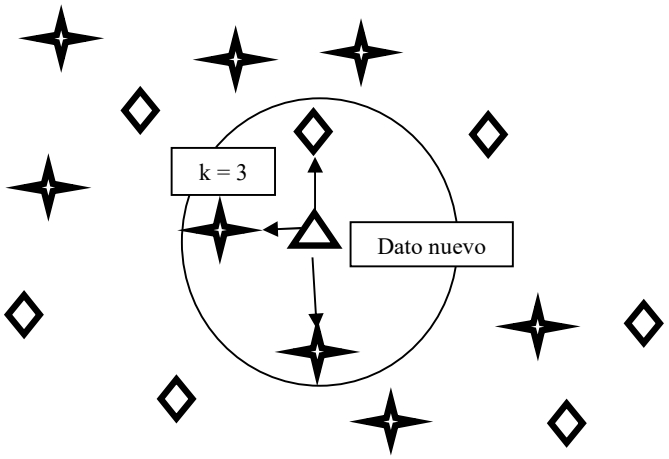
Algoritmo k-NN.

Las técnicas de clasificación k-NN, basan su criterio de aprendizaje en la cultura o adagio popular "Dime con quién andas y te diré quién eres". Dicho de otro modo, los individuos suelen compartir propiedades y características con la comunidad a la que pertenece. De esta manera es posible obtener información de un sujeto al estudiar su vecindad. (Montero, 2009)

A través de un ejemplo se ilustra la forma de operar de este algoritmo. Consta básicamente de tres pasos que serán acompañados de la figura 1-19, que cuenta con 15 instancias y las clases Estrella y Rombo. Como primera medida se calcula la distancia entre el objeto a clasificar (Triangulo) y las instancias del Training Data. Paso seguido, se toman las instancias (para el caso que nos ocupa, k=3) más cercanas según la función implementa.

Por último, se realiza un conteo de los k elegidos teniendo en cuenta la clase con mayor votación, que determina el grupo al que pertenece el objeto a clasificar. Siguiendo el algoritmo para el presente caso, el dato nuevo se corresponde a la clase Estrella (Aprende machine learning , 2018).

Figura 1-19: Algoritmo k-NN.

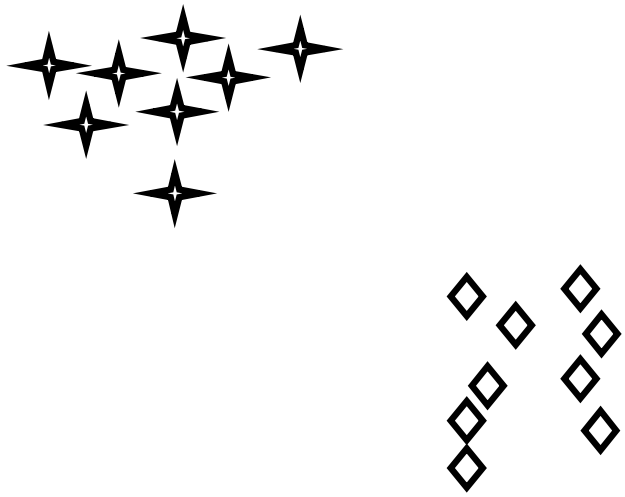


Algoritmo SVM

Las SVM, son modelos eficaces en tareas de clasificación y predicción numérica (regresión), cuyos desarrolladores de la versión actual son Vladimir Vapnik y Corinna Cortés. Este modelo es posible comprenderlo intuitivamente de la siguiente manera: los datos de entrenamiento del modelo son representados en el espacio, (por cuestiones prácticas considerar que el algoritmo cuenta con solo dos clases). Un hiperplano denominado vector es creado, éste separa las clases a dos espacios pronunciados, ante el ingreso de nuevos datos, según los espacios a que pertenece, el modelo realiza la clasificación. (Bedell, 2018).

A continuación, se explica intuitivamente este algoritmo de clasificación. Como primera medida es importante considerar que las SVM, surgen para determinar la forma óptima de clasificación, se cuenta con un Data set de 17 instancias en un espacio bidimensional divididos en las clases “Estrella” y “Rombo”, de 8 y 9 registros respectivamente como se aprecia en la figura 1-20. (Martínez, 2019).

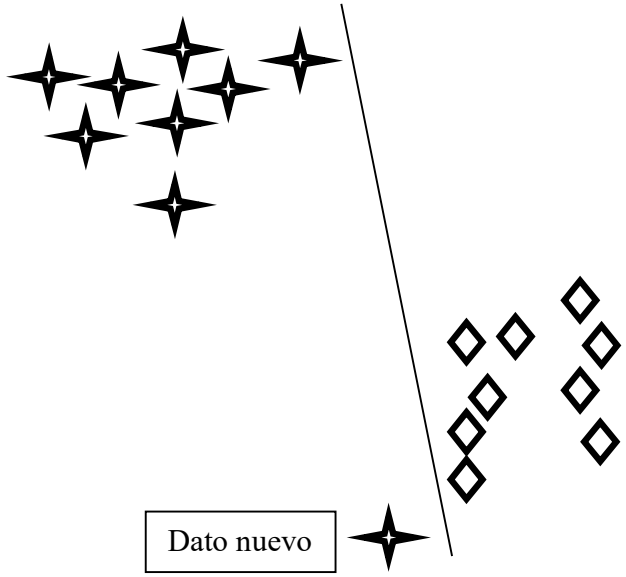
Figura 1-20: Data set, en espacio bidimensional.



Se traza una línea con el objeto de separar las clases, de esta manera ante el ingreso de datos nuevos es posible determinar su clase según su ubicación con respeto a la línea de referencia. A través del método ensayo y error se procederá a realizar unas clasificaciones.

En lo que respecta al caso de la figura 1-21, el modelo predictivo clasifica lo que se encuentra a la izquierda de la línea divisoria como clase “estrella”. Es claro que el dato nuevo no se corresponde con la etiqueta asignada (Martínez, 2019).

Figura 1-21: Caso I. Falla modelo por división incorrecta.



Por lo ilustrado en la figura 1-22, el algoritmo predice que todo dato nuevo ubicado en la parte superior de la línea, debe ser asignado a la clase “estrella”. Por lo que se pude observar, el modelo predictivo falla nuevamente al generalizar. (Martínez, 2019).

Figura 1-22: Caso II. Falla modelo por división incorrecta.

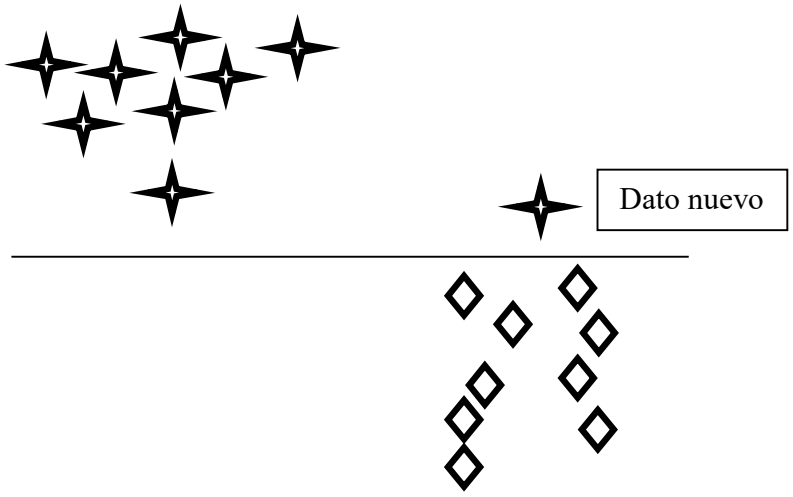
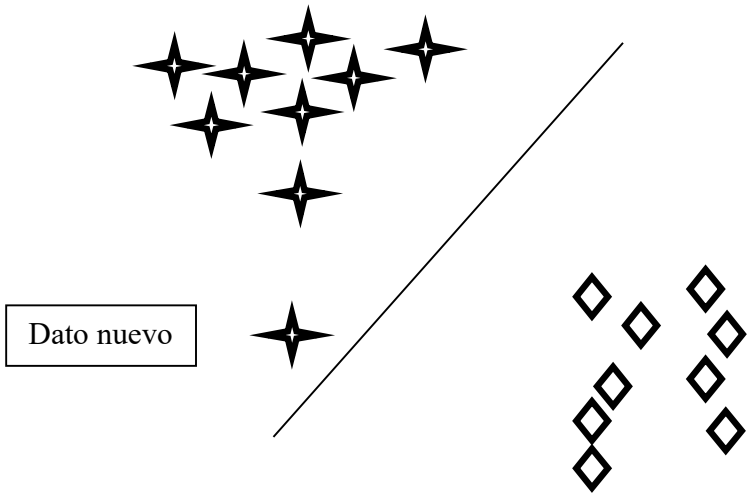


Figura 1-23: Caso III. Modelo con "línea óptima".



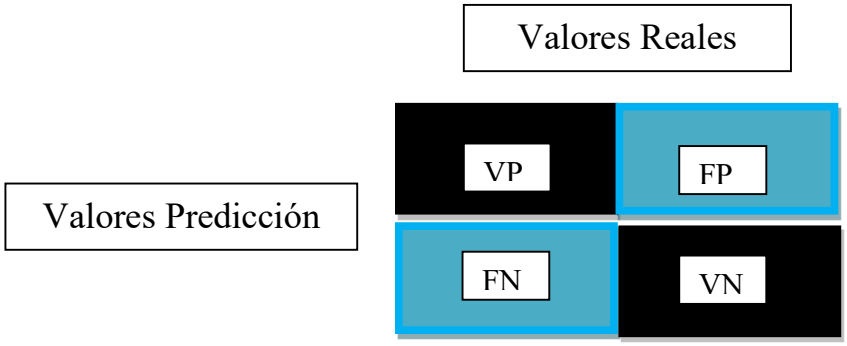
Para determinar correctamente la línea divisoria de clases, esta debe ser trazada de tal manera que maximice el margen entre ambas, como lo expuesto en la figura 1-23. Esta tarea es desempeñada perfectamente por los modelos SVM, donde se cuenta con problemas de tipo multivariable (muchas dimensiones), a diferencia del caso estudiado. De esta manera, se sustituye la denominada "línea óptima" por el hiperplano que se encarga de maximizar el margen de separación entre etiquetas logrando de esta manera una correcta clasificación. (Martínez, 2019).

1.3.2 Métricas de evaluación

Las métricas de evaluación se constituyen como una herramienta para medir el desempeño desde un punto de vista numérico de soluciones propuestas de ML, además suministran un registro del progreso de los distintos modelos para la mejora continua del poder de predicción general antes de su integración en un sistema (Machine Learning En Español.com, 2021) (Singh, 2020)

La matriz de confusión es una técnica para determinar el rendimiento de algoritmos de aprendizaje supervisado en representación matricial. Las filas están constituidas por el número de predicciones de cada clase, por su parte las columnas representan las instancias en la clase real (Serra, 2020).Ver figura 1-24.

Figura 1-24: Matriz de confusión binaria.



Nombre de la fuente: El autor.

Como se aprecia en la imagen se cuenta con dos posibles valores reales y dos posibles valores predichos o de predicción, de esta manera es posible construir la matriz de confusión con los cuatro posibles resultados posibles:

- VP: Valor real positivo. Algoritmo predice un positivo.
- VN: Valor real negativo. Algoritmo predice un negativo.
- FP: Valor real negativo. Algoritmo predice un positivo.
- FN: Valor real positivo. Algoritmo predice un negativo (Barrios, 2019).

A través de un ejemplo se ilustra la comprensión de lo expuesto. Ahora bien, se cuenta con un modelo en ML, para la predicción de pacientes con Covid-19 positivos y negativos, al contar con dos clases se obtiene una matriz de confusión binaria, con las siguientes posibilidades:

1. VP: Paciente con virus. Modelo realiza predicción Covid-19 positivo.
2. VN: Paciente sin virus. Modelo realiza predicción Covid-19 negativo.
3. FP: Paciente sin virus. Modelo realiza predicción Covid-19 positivo.
4. FN: Paciente con virus. Modelo realiza predicción Covid-19 negativo.

Es así como estudiando la figura 1-8, es posible determinar los aciertos del ML en las celdas obscuras. De las posibilidades vistas resultan la exactitud y la precisión, por un lado y la sensibilidad y la especificad por el otro, conocidas como las métricas de la matriz de confusión. (Barrios, 2019).

La exactitud es considerada una de las medidas más implementadas al evaluar el rendimiento de modelos predictivos, definida como la proporción de las predicciones acertadas del algoritmo. (Ver evacuación 1).

$$Exacititud = (VP + VN)/(VP + FP + FN + VN) \quad (1)$$

La precisión o valor predictivo positivo, está definida como el número de predicciones realizadas verdaderamente positivas sobre los resultados positivos (VP y FP). (Ver ecuación 2).

$$Presición = VP/(VP + FP) \quad (2)$$

La sensibilidad o tasa de verdaderos positivos, define el número de instancias de la clase positiva predichas correctamente por el modelo. Se presenta su expresión matemática. (Ver ecuación 3).

$$Sensibilidad = VP/(VP + FN) \quad (3)$$

La especificidad o tasa de verdaderos negativos, define el número de instancias de la

clase negativa predichas correctamente por el modelo. (Ver ecuación 4).

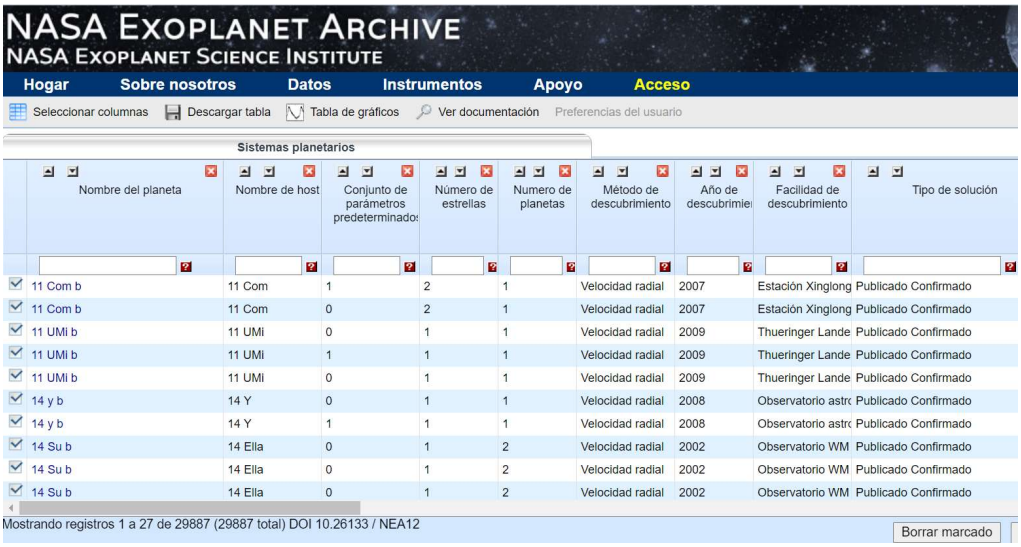
$$Especificidad = VN/(VN + FP) \quad (4)$$

Capítulo 2

2.1 Recolección y filtrado de datos.

La base de datos de exoplanetas para someter a estudio en el presente documento, es un recurso de interés público disponible en el sitio web de la NASA. Esta información es producto de la conquista del espacio realizado a través de distintas fuentes como, los observatorios terrestres y telescopios espaciales mencionados en el capítulo anterior. (NASA, 2021).

Figura 2-1: Enlace para descarga archivo .csv de exoplanetas**.**



Nombre de la fuente: NASA EXOPLANET ARCHIVE. Recuperado de: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

El archivo en formato .csv ha sido descargado de la página web expuesta en la figura 2-1 del vínculo “planetary systems”. Este enlace suministra información concerniente a los exoplanetas confirmados y recursos adicionales que facilitan su estudio y análisis.

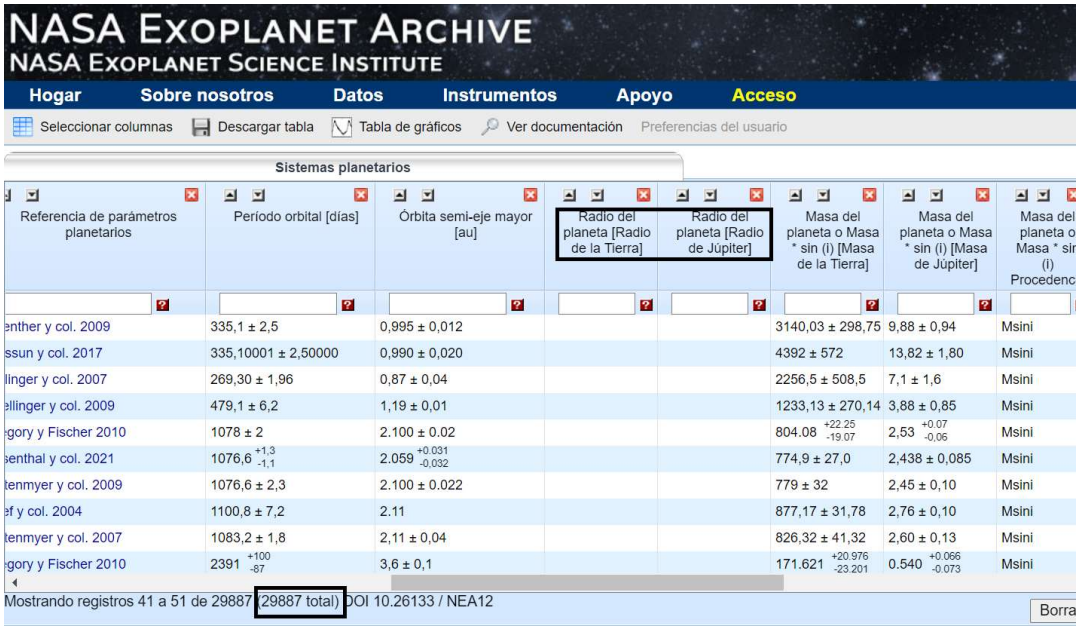
El Data Set dispone de las instancias y atributos obtenidos en las investigaciones, para información adicional consultar la tabla 2-1.

Tabla 2-1: Tabla de datos planetarios extendidos**.**

Instancia y atributos abreviados.	Instancia y atributos completo.	Reseña.
Pl_Name	Planet Name	Nombre con que es reconocido el exoplaneta, en la comunidad científica.
Host_name	Host Name	Estrella sobre la que orbita el exoplaneta en mención.
Default_Flag	Default Parameter Set	Bandera que identifica exoplanetas confirmados
Pl_Orbper	Orbital Period (days)	Tiempo que tarda el exoplaneta en recorrer su órbita dado en días.
Pl_Rade	Planet Radius (Earth Radius)	Línea recta que une el centro del exoplaneta con cualquier punto en la superficie, en unidades radio Tierra.
Pl_Radj	Planet Radios (Júpiter Radios)	Ítem anterior con unidades radio Júpiter.
Pl_Masse	Planet Mass (Earth Mass)	Cantidad de materia del exoplaneta, en unidades masa Tierra.
Pl_Massj	Planet Mass (Júpiter Mass)	Ídem anterior en unidades masa Júpiter.
Pl_Bmasse	Planet Mass *sin(i) (Earth Mass)	La mejor medida de masa en unidades de masa de la Tierra.
Pl_Bmassj	Planet Mass *sin(i) (Jupiter Mass)	La mejor medida de masa en unidades de masa de Júpiter.

Al momento de recolectar la información contenida en la página de la NASA (agosto 29), se contaba con un total de 29.887 cuerpos celestes entre exoplanetas confirmados y no confirmados. A través de un escaneo preliminar se evidencia la ausencia de atributos en instancias, lo que constituye una dificultad a resolver debido a la necesidad de contar con todos los datos pertinentes para la realización de las generalizaciones o predicciones a través de los modelos predictivos generados. Ver figura 2-2.

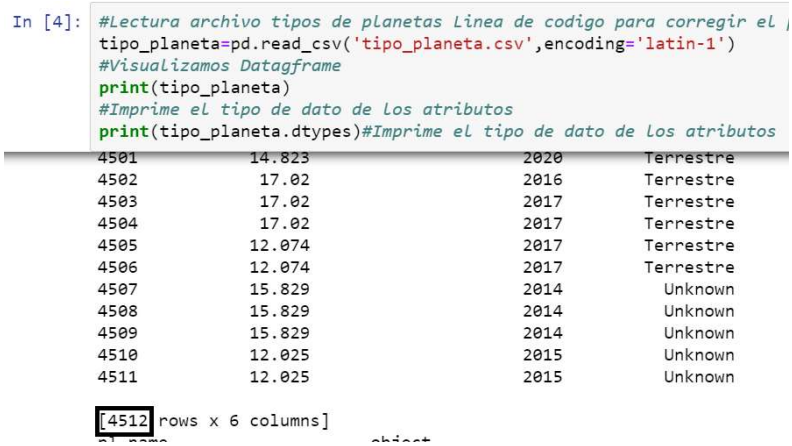
Figura 2-2: Atributos a filtrar y total de cuerpos celestes.



Nombre de la fuente: NASA EXOPLANET ARCHIVE. Recuperado de: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

Es necesario recordar, que el objeto del presente trabajo consiste en la clasificación de los exoplanetas confirmados por la NASA en los distintos tipos de planetas que se conocen: Gigante Gaseoso, Tipo Neptuno, Súper Tierra y Terrestre, a través de los atributos masa del planeta, radio del planeta, período orbital y eje semimayor con técnicas de aprendizaje supervisado. Por lo anterior es necesario tener claro el número de instancias disponibles, con un total de 4512 según se ilustra en la figura 2-3.

Figura 2-3: DataFrame de exoplanetas confirmados.



```
In [4]: #Lectura archivo tipos de planetas Linea de codigo para corregir el
        tipo_planeta=pd.read_csv('tipo_planeta.csv',encoding='latin-1')
        #Visualizamos Datagframe
        print(tipo_planeta)
        #Imprime el tipo de dato de los atributos
        print(tipo_planeta.dtypes)#Imprime el tipo de dato de los atributos
```

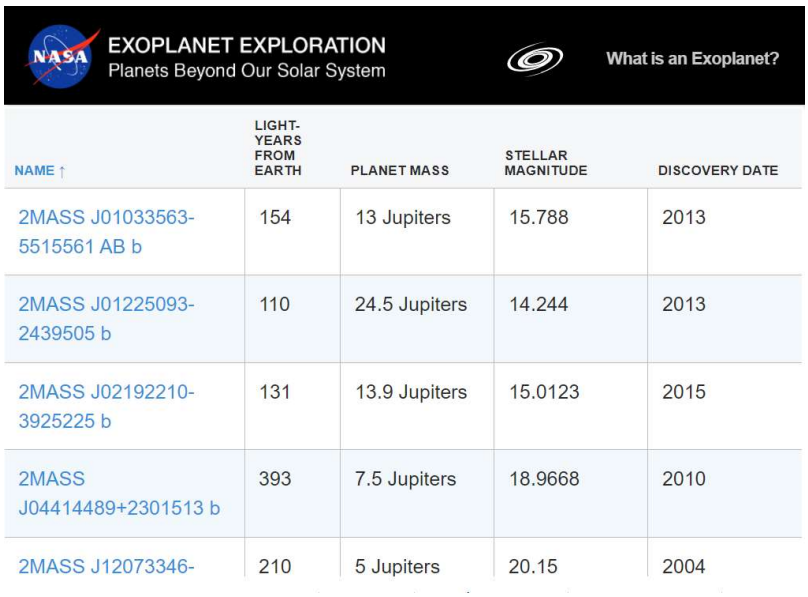
4501	14.823	2020	Terrestre
4502	17.02	2016	Terrestre
4503	17.02	2017	Terrestre
4504	17.02	2017	Terrestre
4505	12.074	2017	Terrestre
4506	12.074	2017	Terrestre
4507	15.829	2014	Unknown
4508	15.829	2014	Unknown
4509	15.829	2014	Unknown
4510	12.025	2015	Unknown
4511	12.025	2015	Unknown

[4512 rows x 6 columns]

Nombre de la fuente: Este trabajo.

Para el desarrollo de la presente investigación, es necesario construir un segundo Data Set proveniente del "catálogo de exoplanetas" de la NASA (EXOPLANET EXPLORATION, 2021). Ver figura 2-4.

Figura 2-4: Segunda base de datos.



EXOPLANET EXPLORATION
Planets Beyond Our Solar System

What is an Exoplanet?

NAME ↑	LIGHT-YEARS FROM EARTH	PLANET MASS	STELLAR MAGNITUDE	DISCOVERY DATE
2MASS J01033563-5515561 AB b	154	13 Jupiters	15.788	2013
2MASS J01225093-2439505 b	110	24.5 Jupiters	14.244	2013
2MASS J02192210-3925225 b	131	13.9 Jupiters	15.0123	2015
2MASS J04414489+2301513 b	393	7.5 Jupiters	18.9668	2010
2MASS J12073346-	210	5 Jupiters	20.15	2004

Nombre de la fuente: Exoplanet Exploration. **Recuperado de:**

https://exoplanets.nasa.gov/discovery/exoplanet-catalog/

Esta base de datos suministra un listado de cuerpos planetarios confirmados en constante actualización, con la primera base de datos. El sitio web contiene la siguiente información del exoplaneta: Name, Light-years from earth, mass, stellar magnitude y discovery date, la tabla 2-2 suministra información adicional.

Tabla 2-2: Descripción catálogo de exoplanetas.

Instancia y atributos	Reseña.
Light years from earth	Distancia del exoplaneta a la Tierra dada en años Luz.
Planet Mass	Cantidad de materia del exoplaneta, en unidades masa Tierra.
Stellar Magnitud	Medida de brillo del cuerpo estelar y cantidad de Luz (Energía).
Discovery Date	Fecha en que se descubre el exoplaneta.

Es así que, con el trabajo realizado hasta el momento, se observa que las bases de datos no suministran la etiqueta del planeta, información importante para realizar la clasificación, por tal razón es necesario adjuntar una columna a la segunda base de datos denominada “Tipo_planeta”. Ver figura 2-5.

Figura 2-5: Archivo con columna de clases.

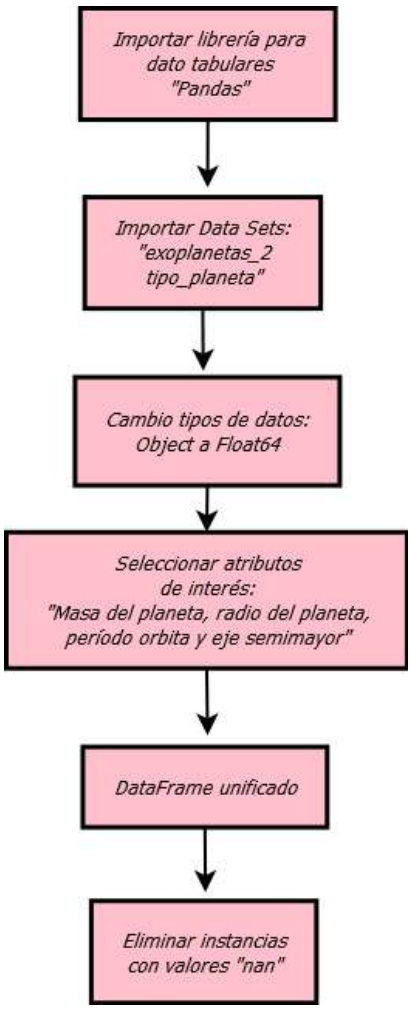
pl_name	A AÑOS LUZ	Valor_masa_	MAGNITUD E	mj_mt	FECHA DE DE	Tipo_planeta
11 Comae Be	304	19.4	472.307	317.83	2007	Gigante gaseoso
11 Ursae Mir	409	14.74	5.013	317.83	2009	Gigante gaseoso
14 Andróme	246	4.8	523.133	317.83	2008	Gigante gaseoso
14 Herculis b	58	4.66	661.935	317.83	2002	Gigante gaseoso
16 Cygni B b	69	1.78	6.215	317.83	1996	Gigante gaseoso
18 Delphini b	249	10.3	551.048	317.83	2008	Gigante gaseoso
1RXS J160929	454	8	12.618	317.83	2008	Gigante gaseoso
24 Bootis b	313	0.91	5.59	317.83	2018	Gigante gaseoso
24 Sextantis	235	1.99	64.535	317.83	2010	Gigante gaseoso
24 Sextantis	235	0.86	64.535	317.83	2010	Gigante gaseoso
2MASS J0103	154	13	15.788	317.83	2013	Gigante gaseoso
2MASS J0122	110	24.5	14.244	317.83	2013	Gigante gaseoso
2MASS J0219	131	13.9	150.123	317.83	2015	Gigante gaseoso
2MASS J0441	393	7.5	189.668	317.83	2010	Gigante gaseoso
2MASS J1207	210	5	20.15	317.83	2004	Gigante gaseoso
2MASS J1938	1293	1.9	12.651	317.83	2015	Gigante gaseoso
2MASS J2140	108	20.95	173.961	317.83	2009	Gigante gaseoso
2MASS J2236	227	12.5	12.368	317.83	2016	Gigante gaseoso
30 Arietis B b	146	13.82	709.209	317.83	2009	Gigante gaseoso
42 Draconis b	296	3.88	48.262	317.83	2008	Gigante gaseoso

Nombre de la fuente: Este trabajo.

1.2 Pre-procesamiento.

La calidad de los datos con que se alimenta un modelo predictivo como los implementados en este trabajo, determina en gran medida el rendimiento del modelo, razón por la cual, es necesario pre-procesarlos, para tratar situaciones frecuentes en los Data Set como la perdida de datos causado por medidas no aplicables, espacio en blanco de encuestas e incluso error humano. Para lograr lo anterior, en Python se aplican los pasos expuestos en el diagrama de flujo de la figura 2-6. (Roman, 2019).

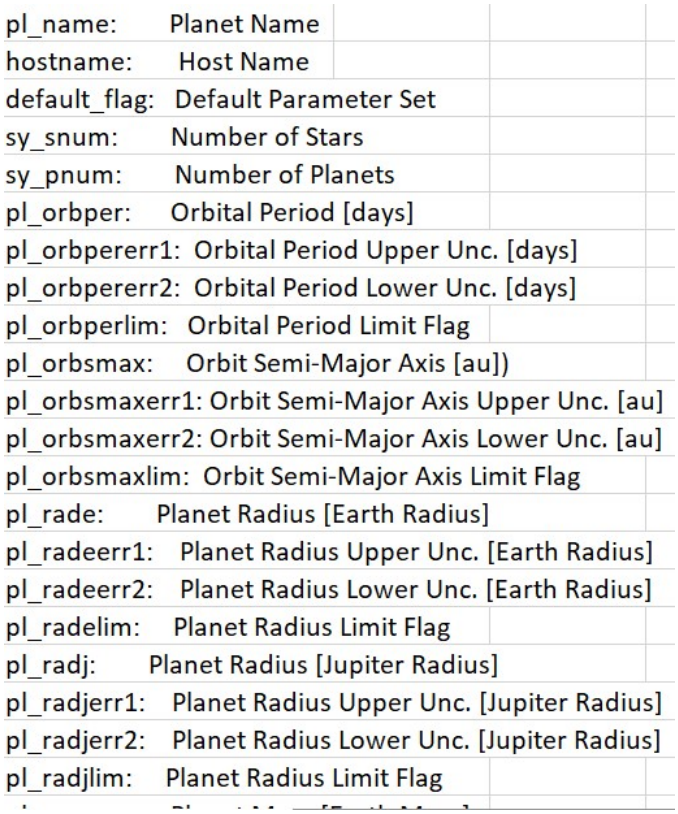
Figura 2-6: Diagrama de flujo pre-procesamiento de datos.



Nombre de la fuente: El autor.

Como se había mencionado con anterioridad, se cuenta con los Data Sets de las ilustraciones 2-2 y 2-5. Cada una contiene sus instancias y atributos producto de la investigación realizada por los diversos observatorios ya mencionados. Por consiguiente, es necesario recurrir a la información adicional disponible en el archivo .csv, para tener pleno conocimiento de la naturaleza de cada dato. Ver figura 2.7.

Figura 2-7: Información de encabezado. Referencia de cada atributo.



Nombre de la fuente: NASA EXOPLANET ARCHIVE. Recuperado de: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

Ahora bien, las gráficas son potentes herramientas en la visualización de datos, que permiten exponer amigablemente información compleja, además, de adaptar este conocimiento al público que va dirigido (alumnos, docentes, investigadores, divulgadores) (UNIR, 2020).

Por lo anterior, los datos adquiridos en las bases de datos requieren ser estandarizados para su visualización en gráficas de dispersión como se verá más adelante.

El sitio web de la NASA dispone de un flag denominado "Conjunto de parámetros determinados". A través de éste, es posible realizar la descarga del archivo exoplanetas_2.cvs con dos modalidades. El parámetro por defecto "0" suministra los datos de los exoplanetas confirmados y no confirmados. Como este trabajo de investigación requiere la implementación de planetas confirmados es necesario descargar el archivo con el parámetro en "1" para descartar aquellos datos de exoplanetas que no han sido confirmados hasta el momento. Ver figura 2-8.

Figura 2-8: Descarga archivo con parámetro en "1".

Nombre del planeta	Nombre de host	Conjunto de parámetros predeterminados	Número de estrellas	Numero de planetas
		1		
11 Com b	11 Com	1	2	1
11 UMi b	11 UMi	1	1	1
14 y b	14 Y	1	1	1
14 Su b	14 Ella	1	1	2
16 Cyg B b	16 Cyg B	1	3	1
18 Del b	18 Del	1	2	1
1RXS J160929.1-210524 b	1RXS J160929.1-	1	1	1
24 Boo b	24 Boo	1	1	1
24 Sexo b	24 sexo	1	1	2
24 Sexo c	24 sexo	1	1	2
2MASS J01033563-5515561 AB b	2MASS J0103356	1	2	1
2MASS J01225093-2439505 b	2MASS J0122509	1	1	1

Nombre de la fuente: NASA EXOPLANET ARCHIVE. Recuperado de: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

En un apartado anterior se manifestó la ausencia de datos en los atributos de algunos exoplanetas observados. Python a través del DataFrame los representa con la abreviación "nan" o valor que no es número (del inglés Not a Number). (DelftStack, 2020). Estudiando las bases de datos (en especial el archivo exoplanetas_2.csv), es claro

que esta situación se evidencia con mayor frecuencia en columnas que no resultan de interés en el presente trabajo.

Es importante resaltar que como atributos de interés se toman: período orbital, radio del planeta, eje semimayor disponibles en la primera base de datos y masa del planeta suministrada por el segundo recurso de la NASA.

De esta manera, del primer archivo se conservan los mencionados atributos eliminando las columnas restantes, por otra parte, del segundo archivo se conserva la información en su totalidad junto a la columna tipos de planeta, que indica las etiquetas de clasificación de los modelos predictivos.

Figura 2-9: DataFrame con exoplanetas disponibles.

```
#Visualizamos Datagframe definitivo
print(tabla_definitiva)
```

```
4437     692.86940  Gigante gaseoso
4439    1875.19700  Gigante gaseoso
4440     200.23290  Gigante gaseoso
4441     425.89220  Gigante gaseoso
4442     255.85315  Gigante gaseoso
4443     467.21010  Gigante gaseoso
4447     152.55840  Gigante gaseoso
4449     293.03926  Gigante gaseoso
4452    4386.05400  Gigante gaseoso
4454    1274.49830  Gigante gaseoso
4455     355.96960  Gigante gaseoso
4456     188.15536  Gigante gaseoso
4457     232.33373  Gigante gaseoso
4461       4.78000     Super Tierra
4463     179.89178  Gigante gaseoso
4468     378.21770  Gigante gaseoso
4469    1398.45200  Gigante gaseoso
4470     225.34147  Gigante gaseoso

[1672 rows x 6 columns]
```

Nombre de la fuente: Este trabajo.

Se procede a unificar las tablas realizando un último filtrado de las instancias donde se encuentren valores “nan”. Tras todo lo anterior se obtiene el DataFrame “tabla_definitva” con un total de 1672 exoplanetas y los atributos que se ilustran en las figuras 2-9 y 2-10.

Figura 2-10: Atributos de interés.

```
#Visualizamos Datagframe definitivo
print(tabla_definitiva)
```

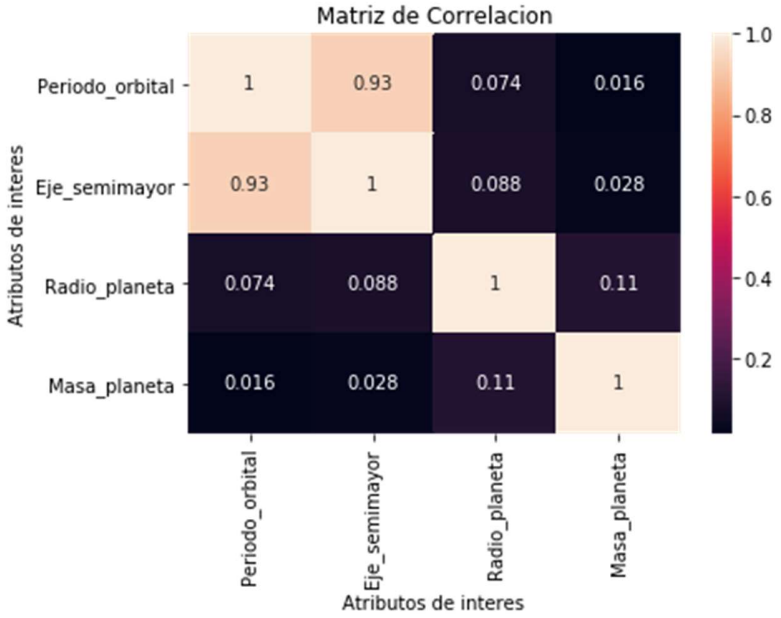
	Nombre_planeta	Periodo_orbital	Eje_semimayor	Radio_planeta	\
70	CFHTWIR-Oph 98 b	8.040000e+06	200.00000	20.849	
76	CoRoT-10 b	1.324060e+01	0.10550	10.870	
77	CoRoT-11 b	2.994330e+00	0.04360	16.030	
78	CoRoT-12 b	2.828042e+00	0.04016	16.140	
79	CoRoT-13 b	4.035190e+00	0.05100	9.920	
80	CoRoT-14 b	1.512140e+00	0.02700	12.220	
81	CoRoT-16 b	5.352270e+00	0.06180	13.110	
82	CoRoT-17 b	3.768100e+00	0.04610	11.430	
83	CoRoT-18 b	1.900069e+00	0.02950	14.680	
84	CoRoT-19 b	3.897130e+00	0.05180	14.460	
85	CoRoT-2 b	1.742994e+00	0.02798	16.432	
88	CoRoT-21 b	2.724740e+00	0.04170	14.572	
89	CoRoT-22 b	9.755980e+00	0.09200	4.880	
90	CoRoT-23 b	3.631300e+00	0.04800	11.770	
91	CoRoT-24 b	5.113400e+00	0.05600	3.700	
92	CoRoT-24 c	1.175900e+01	0.09800	5.000	
93	CoRoT-25 b	4.860690e+00	0.05780	12.110	
94	CoRoT-26 b	4.204740e+00	0.05260	14.120	

Nombre de la fuente: Este trabajo.

Al realizar el análisis de información sobre un Data Set, se cuenta con una herramienta muy valiosa que son las correlaciones. Esta técnica tiene base estadística y por lo tanto matemática, con las cuales es posible determinar la fuerza y el sentido de la relación entre variables. (Suárez, 2015)

Python suministra esta herramienta de estadística descriptiva para datos multivariables llamada Matriz de correlación, esta representa las “correlaciones” entre pares de variables en un dato dado, representados en cada columna y fila, donde los valores de la matriz corresponden al coeficiente de correlación entre variables. Se procede a realizar este estudio con los atributos de interés filtrados de la figura 2-10.

Figura 2-11: Matriz de correlación.



Nombre de la fuente: Este trabajo.

Realizando un análisis a la matriz de correlación de la figura 2-11, se observa que el par "Período orbital" "Eje semimayor" es apropiado para ser implementado en algoritmos de aprendizaje, teniendo una correlación cercana a '1', siendo '1' la mayor correlación que puede existir entre los atributos a estudiar.

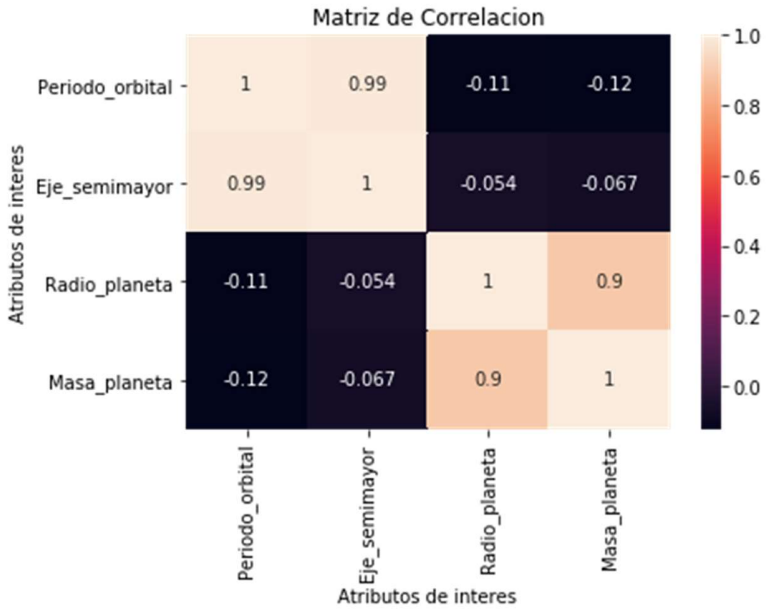
Figura 2-12: Pares ordenados.

```
Periodo_orbital  Masa_planeta       0.015706
Masa_planeta     Periodo_orbital    0.015706
Eje_semimayor    Masa_planeta       0.028415
Masa_planeta     Eje_semimayor      0.028415
Periodo_orbital  Radio_planeta      0.074044
Radio_planeta    Periodo_orbital    0.074044
Eje_semimayor    Radio_planeta      0.087791
Radio_planeta    Eje_semimayor      0.087791
                 Masa_planeta       0.109059
Masa_planeta     Radio_planeta      0.109059
Periodo_orbital  Eje_semimayor      0.934083
Eje_semimayor    Periodo_orbital    0.934083
Periodo_orbital  Periodo_orbital    1.000000
Eje_semimayor    Eje_semimayor      1.000000
Radio_planeta    Radio_planeta      1.000000
Masa_planeta     Masa_planeta       1.000000
dtype: float64
```

Nombre de la fuente: Este trabajo.

En la ilustración 2-12, se pueden observar los valores de pares ordenados de las correlaciones entre los atributos de interés, donde se confirma un valor de 0.9340883 de correlación existente entre los atributos “Eje semimayor” y “Período orbital”. Con el objeto de lograr mejores resultados, se continua el estudio, pero ahora haciendo uso de la base de datos en escala logarítmica.

Figura 2-13: Matriz de correlación en representación logarítmica.



Nombre de la fuente: Este trabajo.

De la matriz de correlación expuesta en la figura 2-13, se observa una mejor respuesta al implementar los datos en este tipo de escala, obteniendo valores superiores a 0.8. Con el apoyo de la ilustración 2-14, la cual contiene los valores de pares ordenados, se puede confirmar un coeficiente de correlación de 0.898544 correspondiente a “Masa_planeta” “Radio_planeta” y 0.988243 con “Período orbital” “Eje semimayor”.

Figura 2-14: Pares ordenados en representación logarítmica.

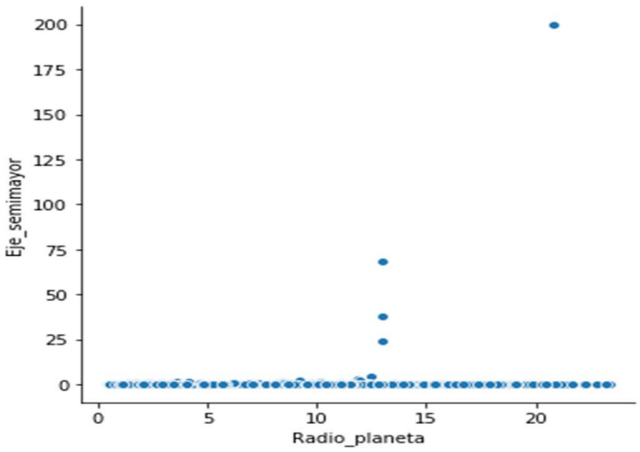
```
Periodo_orbital  Masa_planeta      -0.122809
Masa_planeta     Periodo_orbital   -0.122809
Periodo_orbital  Radio_planeta     -0.114323
Radio_planeta    Periodo_orbital   -0.114323
Eje_semimayor    Masa_planeta      -0.066925
Masa_planeta     Eje_semimayor     -0.066925
Eje_semimayor    Radio_planeta     -0.053881
Radio_planeta    Eje_semimayor     -0.053881
                 Masa_planeta       0.898544
Masa_planeta     Radio_planeta      0.898544
Periodo_orbital  Eje_semimayor      0.988243
Eje_semimayor    Periodo_orbital    0.988243
Periodo_orbital  Periodo_orbital    1.000000
Eje_semimayor    Eje_semimayor      1.000000
Radio_planeta    Radio_planeta      1.000000
Masa_planeta     Masa_planeta       1.000000
dtype: float64
```

Nombre de la fuente: Este trabajo.

Las gráficas se constituyen en un recurso de gran ayuda en la solución de problemas de Ingeniería. Es posible su construcción a través de algunas instrucciones en Python con el apoyo de librerías como Matplotlib y Seaborn. (Metodos numericos , 2021)

En la figura 2-15 se inicia con la visualización de los atributos “Radio del planeta” y “Eje semimayor”. Se puede observar que la representación suministrada por defecto en Python no permite deducir información relevante, lo que obliga a definir los límites de la gráfica.

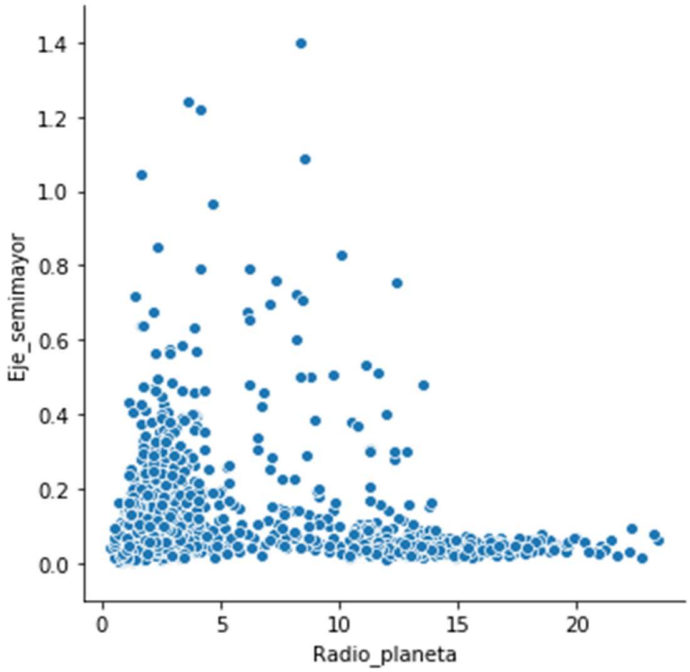
Figura 2-15: Gráfica con límites por defecto.



Nombre de la fuente: Este trabajo.

Implementado las instrucciones necesarias para definir el límite de la coordenada vertical, se logra enriquecer la gráfica al imprimir los datos en un rango más apropiado para su visualización. Aun así, es evidente que en la base de datos se presenta un número elevado de exoplanetas con eje semimayor inferior a 0.4 que disminuye a medida que este valor se aproxima a 1.4. Lo anterior conlleva a una visualización con espaciado irregular por la densidad de información presente en una sección de la gráfica, esta situación puede ser observada en la ilustración 2-16.

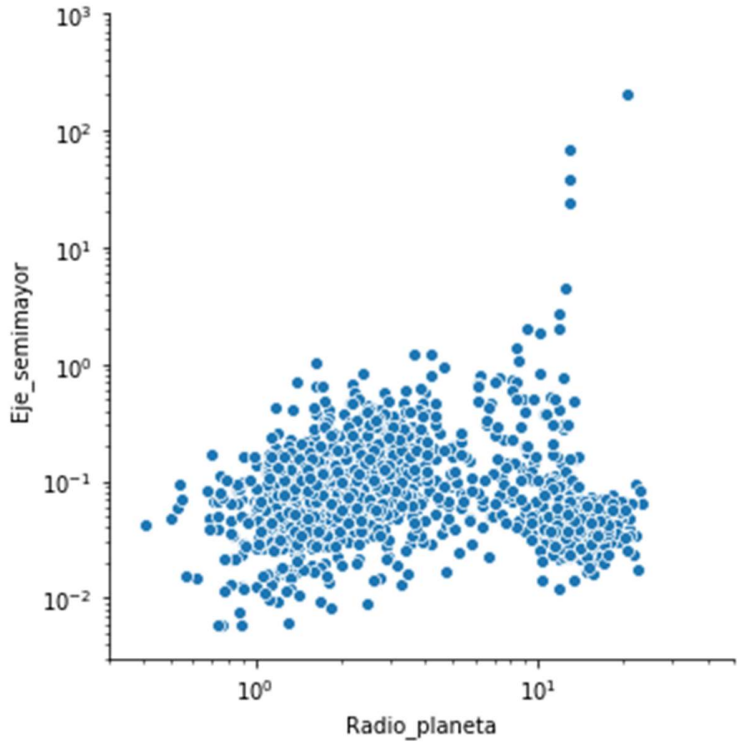
Figura 2-16: Se definen límites de ordenada.



Nombre de la fuente: Este trabajo.

Para solucionar la situación es posible hacer uso de escalas logarítmicas, a través de ellas se realiza la conversión del grafico donde se relacionan las variables, con el agregado de representar valores de magnitudes muy diferentes, situación que se ajusta perfectamente al presente caso. Gracias al trabajo ejecutado, la gráfica despliega la información contenida en la base de datos en su totalidad. Ver figura 2-17. (Sanchez, 2021).

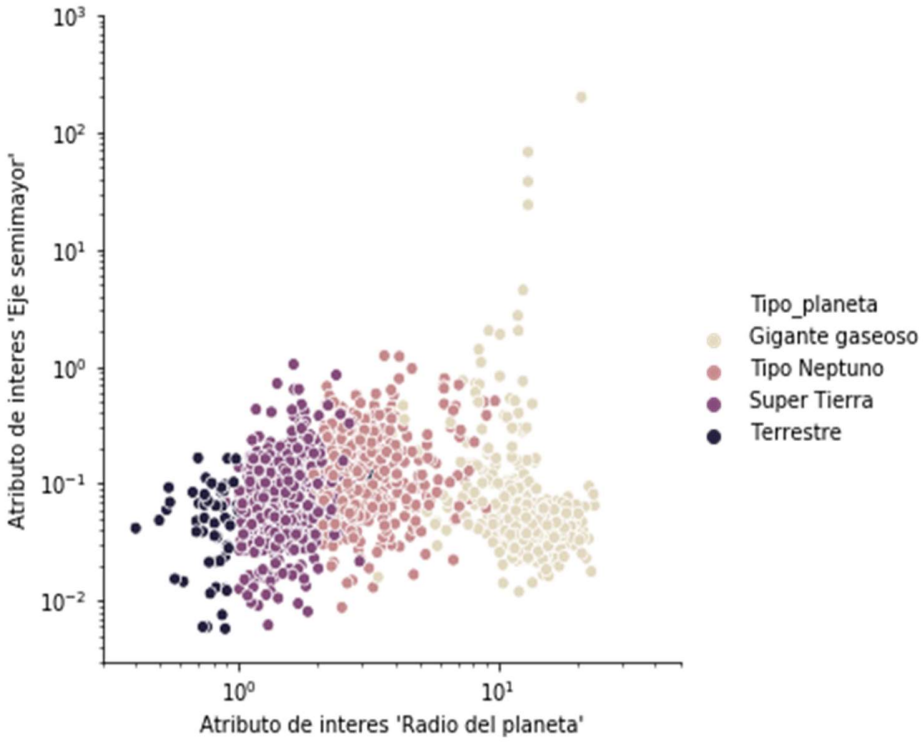
Figura 2-17: Abscisa y ordenada en escala logarítmica.



Nombre de la fuente: Este trabajo.

En la información desplegada no es posible apreciar la clasificación contenida en la columna "Tipo_planeta" que constituye las clases de los modelos predictivos. La ilustración 2-18 es una gráfica de dispersión, donde se relacionan los atributos de interés "radio del planeta" "eje semimayor" y la leyenda de etiquetas. Como primera medida, es posible confirmar lo expresado en la matriz de correlación de la figura 2-13 y los pares ordenados de la ilustración 2-14 donde se presentaba un coeficiente con valor de -0.053881, lo que indica muy baja correlación. Además, en cada clase de Tipos de planeta se observa una pequeña presencia de las demás etiquetas.

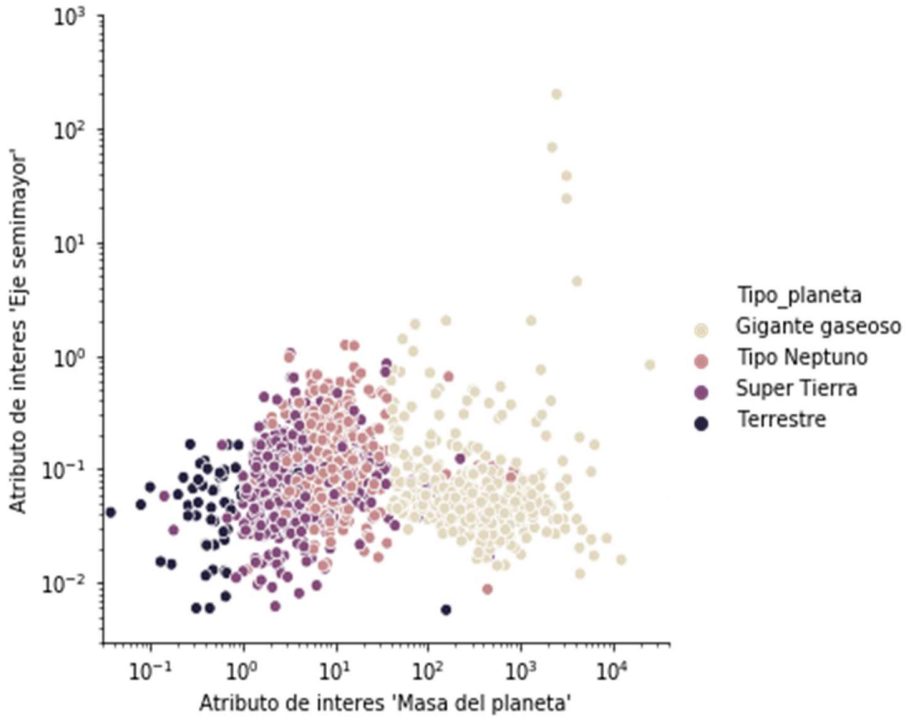
Figura 2-18: Gráfica de dispersión 1.



Nombre de la fuente: Este trabajo.

En la gráfica de dispersión de la figura 2-19 se puede observar un patrón similar al estudio anterior (datos solapados). Gracias al análisis matemático suministrado por la ilustración 2-14, se observa una menor correlación de datos (coeficiente de correlación -0.066925), por lo que se continúa con el estudio de gráficas.

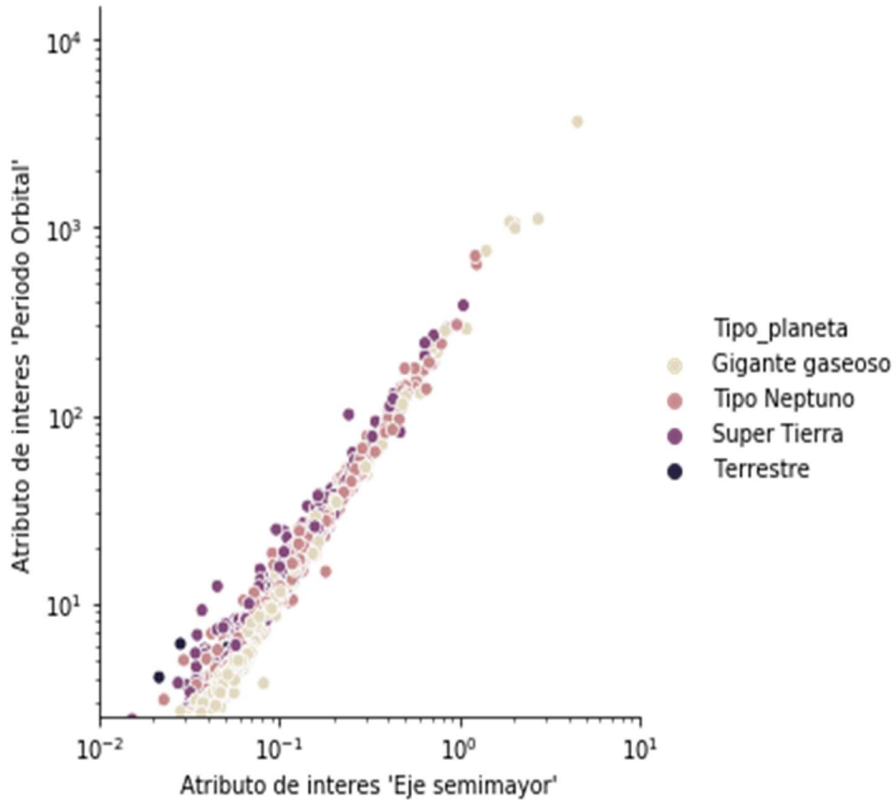
Figura 2-19: Gráfica de dispersión 2.



Nombre de la fuente: Este trabajo.

A través la de gráfica 2-20 se observa la mejor correlación de datos obtenida, donde se puede observar que las dos variables aumentan al mismo tiempo indicando una correlación positiva (Minitab, 2020). Para corroborar lo anterior, se consulta una vez más los valores de pares ordenados con un coeficiente de correlación de 0.988243.

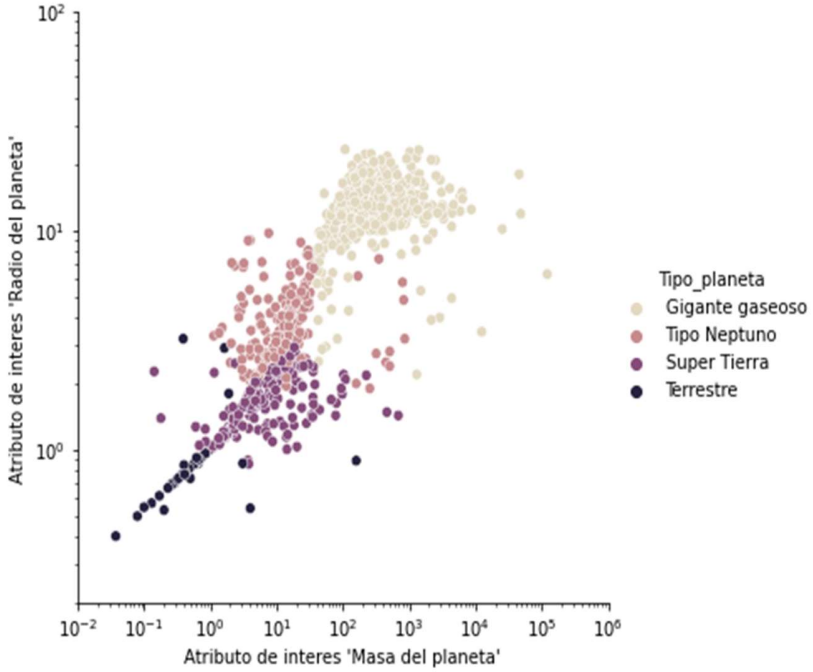
Figura 2-20: Gráfica de dispersión 3.



Nombre de la fuente: Este trabajo.

Al graficar los atributos “Masa del planeta” y “Radio del planeta” en la ilustración 2-21 se observa correlación, los datos se encuentran en escala logarítmica con un valor de correlación de 0.898544 como se observa en la gráfica 2-14, adicionalmente se cuenta con un marcado agrupamiento.

Figura 2-21: Gráfica de dispersión 4.



Nombre de la fuente: Este trabajo.

Es importante tener en cuenta que, al implementar técnicas de aprendizaje automático, es recomendable incluir variables con alta correlación ya que estas tienen el mayor poder predictivo del modelo. Esta situación se cumple con los atributos de la gráfica 2-20 (Aws, 2021). Adicionalmente las clases de la base de datos deben presentar agrupamiento, condición que se cumple por lo observado en las ilustraciones 2-18, 2-19 y 2-21.

A continuación, se procede a sinterizar la información estudiada, para ilustrar la metodología al implementar modelos predictivos de aprendizaje supervisado. Como primera medida se da inicio a la recolección de datos, proceso explorado en la sección 2.1 donde se realizó la descarga de las dos bases de datos sometiéndola a un escaneo general y filtrado inicial de datos nulos. En el pre-procesamiento visto en la sección 2.2 se procedió a construir un DataFrame unificado al que se adiciona la columna de etiquetas. Esta estructura de datos es sometida a un estudio de matrices de correlación y por último se visualizan los atributos con la leyenda de etiquetas a través de las gráficas de dispersión.

De esta manera, se da inicio al modelado de los algoritmos de aprendizaje supervisado en Python. Como en todo código, es necesario importar las librerías creadas por terceros que facilitan la programación al contener las funciones puntales en la resolución de problemas. En la figura 2-22 se realiza este proceso con la librería Scikit-Learn que ofrece los módulos y algoritmos para el aprendizaje y trabajo científico de datos (Platzi, 2018) (Universidad de Alcalá., 2021)

Figura 2-22: Librería Scikit-Learn.

```
########## APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING ##########

from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
```

Nombre de la fuente: Este trabajo.

Se cuenta con una base de datos unificada como se había mencionado con anterioridad donde están contenidos los atributos y etiquetas de clasificación, por lo que se procede a construir dos DataFrame, una que contenga las características y otra las clases. A continuación, se crean los dos subconjuntos necesarios para los procesos de entrenamiento (requerido en la construcción del modelo) y prueba (datos desconocidos para evaluar desempeño), destinando un 80 y 20% de la base de datos respectivamente, de no realizar lo anterior, el algoritmo tan solo logra un ajuste perfecto a los datos de

entrenamiento (memoriza no aprende) lo que conlleva a malas predicciones (Bit Degree, 2020) (Muller & Guido, 2017) (Singh, 2020).Ver figura 2-23.

Figura 2-23: Subconjuntos “entrenamiento” y “prueba”.

```
##########################################################################
# MODELOS CON ATRIBUTOS RADIO-MASA-PERIODO_ORBITAL-EJE_SEMIMAYOR DEL PLANETA #
############## DATOS SIN COVERSION A LOGARITMO MATURAL ####################
##########################################################################

X_radio_masa_orbital= tabla_definitiva[['Radio_planeta','Masa_planeta','Periodo_orbital','Eje_semimayor']]
y_radio_masa_orbital = tabla_definitiva[['Tipo_planeta']]
#Verifico caracteristica y etiquetas
print(X_radio_masa_orbital)
print(y_radio_masa_orbital)

#Ahora procedemos a separar los datos de entrenamiento y prueba para proceder a construir los modelos.
X_train_total, X_test_total, y_train_total, y_test_total = train_test_split(X_radio_masa_orbital, y_radio_masa_orbital,
print('Son {} datos para entrenamiento y {} datos para prueba'.format(X_train_total.shape[0], X_test_total.shape[0]))
```

Nombre de la fuente: Este trabajo.

Muchos algoritmos de aprendizaje automático usados en ciencias de datos, requieren normalizar los datos antes de entrenar el modelo, este proceso comprime o extiende los valores delimitando su rango, mejorando el funcionamiento de los modelos. En la figura 2-24 se ejecuta el programa respectivo en la consola interactiva de Jupyter. (Morante, 2018).

Figura 2-24: Normalizar datos.

```
# El escalador de objetos (modelo)
scaler = StandardScaler ()
# Ajustar y transformar los datos
X_train_parcial = scaler.fit_transform (X_train_parcial)
X_test_parcial = scaler.fit_transform (X_test_parcial)
```

Nombre de la fuente: Este trabajo.

Es el momento de desarrollar el modelo propiamente dicho. En la figura 2-25 se puede estudiar la estructura general, donde la primera línea de código es modificada según el algoritmo a implementar, la segunda instrucción realiza el entrenamiento con el Training Data (atributos y etiquetas), por último, se procede a realizar una predicción con los datos de prueba (características no vistas por el modelo). (aprendeIA, 2021)

Figura 2-25: Líneas de código, desarrollo ML.

```
#Modelo de Vecinos más Cercanos
algoritmo_knn = KNeighborsClassifier(n_neighbors=7)
algoritmo_knn.fit(X_train_total, y_train_total)
Y_pred_knn = algoritmo_knn.predict(X_test_total)
```

Nombre de la fuente: Este trabajo.

Capítulo 3

3.1 Análisis de Resultados.

Como se mencionó con anterioridad, a través del presente trabajo se implementan los algoritmos de aprendizaje supervisado SVM, k-NN y Árbol de decisiones a la base de datos de la NASA de exoplanetas con los cuatro atributos masa del planeta, radio del planeta, período orbital y eje semimayor. Los cuerpos celestes serán clasificados en las etiquetas Súper Tierra, Terrestre, Tipo Neptuno y Gigante Gaseoso.

Para ello cada algoritmo es modelado cuatro veces con las siguientes consideraciones: en los primeros se implementan las características masa del planeta y radio del planeta, para los segundos modelos, los atributos escogidos son período orbital y eje semimayor, posteriormente son sustituidos por masa del planeta, radio del planeta y período orbital, como paso finál se implementan los algoritmos con todas las características.

Por lo dicho, se inicia el análisis con los atributos masa y radio del planeta. En la figura 3-1 se tienen los resultados de precisión y exactitud con los datos de características en escala normal. Los algoritmos SVM y k-NN son los modelos con mayor rendimiento frente al de árbol de decisiones, observando el gráfico de dispersión de la figura 2-21 se puede concluir que gracias a la correlación y el agrupamiento obtenido se logran estas métricas.

Figura 3-1: Métricas con masa y radio del planeta.

Modelos ML(Datos originales)	Precisión Masa-Radio	Exactitud Masa-Radio
Máquinas de Vectores	0.901493	0.901493
k-Vecinos más cercanos	0.907463	0.907463
Árbol de decisiones	0.635821	0.635821

Nombre de la fuente: Este trabajo.

La ilustración 3-2 corresponde a la matriz de confusión del algoritmo SVM. En esta representación matricial, Python asigna una etiqueta de clasificación a cada una de las filas de la siguiente manera: fila “0” corresponde a Gigante Gaseoso, “1” Tipo Neptuno, “2” Terrestre y “3” Súper Tierra. De esta manera el algoritmo predijo correctamente todos los exoplanetas Gigantes Gaseoso, de los datos Tipo Neptuno clasificó incorrectamente 3 como Súper Tierra, en cuanto a las instancias Terrestre generalizó perfectamente y por último de los datos de prueba con clase Súper Tierra clasifica incorrectamente 15 como Tipo Neptuno.

En la diagonal principal se tienen los verdaderos positivos predichos, quedando en las demás celdas los falsos positivos, es posible validar la precisión observada en la figura 3-1 con el apoyo de la ecuación 2.

$$Presición = VP/(VP + FP) \tag{5}$$

$$Presición = (90 + 110 + 102)/(90 + 110 + 102 + 11 + 15 + 3 + 4) \tag{6}$$

$$Presición = 0.901492 \tag{7}$$

Figura 3-2: Matriz de confusión SVM.

	0	1	2	3
0	90	0	0	4
1	0	110	0	3
2	0	11	0	0
3	0	15	0	102

Nombre de la fuente: Este trabajo.

Continuando el estudio, se toman las características período orbital y eje semimayor, obteniendo las métricas de la ilustración 3-3. Gracias al consolidado de la figura 3-4 se observa que el rendimiento de los algoritmos disminuye en más del 50% con respecto a los modelos con atributos masa del planeta y radio del planeta, lo que con cuerda con lo observado en la sección 2 figura 2-20, donde a pesar de que el gráfico de dispersión demostraba una muy buena correlación se presenta solapamiento de datos, generando un

impacto negativo en el desempeño de los modelos.

Figura 3-3: Métricas con Período orbital y eje semimayor.

Modelos ML(Datos originales)	Precisión Orbital-Eje	Exactitud Orbital-Eje
Máquinas de Vectores	0.346269	0.346269
k-Vecinos más cercanos	0.343284	0.343284
Árbol de decisiones	0.080597	0.080597

Nombre de la fuente: Este trabajo.

En la figura 3-4 se sintetiza el rendimiento de los modelos implementados al momento.

Figura 3-4: Métricas comparativas primeros modelos.

Modelos ML(Datos originales)	Precisión Masa-Radio	Precisión Orbital-Eje	Exactitud Masa-Radio	Exactitud Orbital-Eje
Máquinas de Vectores	0.901493	0.346269	0.901493	0.346269
k-Vecinos más cercanos	0.907463	0.343284	0.907463	0.343284
Árbol de decisiones	0.635821	0.080597	0.635821	0.080597

Nombre de la fuente: Este trabajo.

Como paso adicional se realizan los modelos predictivos con las características masa del planeta, radio del planeta y período orbital obteniendo las métricas de la ilustración 3-5. Comparando este resultado con lo observado en la figura 3-4, no se presenta mejora en el rendimiento de los modelos predictivos SVM y k-NN, con un incremento de 10 unidades porcentuales en el algoritmo árbol de decisiones.

Figura 3-5: Métricas con radio, masa y periodo orbital.

Modelos ML (Datos originales)	Precisión radio-masa-orbital	Exactitud radio-masa-orbital
Máquinas de Vectores	0.901493	0.901493
k-Vecinos más cercanos	0.907463	0.907463
Árbol de decisiones	0.731343	0.731343

Nombre de la fuente: Este trabajo.

A continuación, se incluyen todas las características con el resultado de la ilustración 3-6. En el modelo predictivo SMV se mantienen constantes la precisión y exactitud,

incrementándose en una unidad porcentual el algoritmo k-NN y un decremento de tres el árbol de decisiones, de esta manera, el atributo eje semimayor no genera un cambio significativo en el rendimiento de los modelos, lo que corrobora una vez el resultado del solapamiento de datos con las métricas período orbital y eje semimayor.

Figura 3-6: Métricas con masa, radio, período orbital y eje semimayor.

Modelos ML (datos originales)	Precisión Total	Exactitud Total
Máquinas de Vectores	0.901493	0.901493
k-Vecinos más cercanos	0.910448	0.910448
Árbol de decisiones	0.707463	0.707463

Nombre de la fuente: Este trabajo.

La figura 3-7 sintetiza el rendimiento de los segundos modelos implementados al momento. En los resultados parciales se realiza el entrenamiento de los algoritmos con las características masa del planeta, radio del planeta y período orbital, en las métricas "Total" se incluye todos los atributos.

Figura 3-7: Métricas comparativas segundos modelos.

Modelos ML (datos originales)	Precisión Parcial	Precisión Total	Exactitud Parcial	Exactitud Total
Máquinas de Vectores	0.901493	0.901493	0.901493	0.901493
k-Vecinos más cercanos	0.907463	0.910448	0.907463	0.910448
Árbol de decisiones	0.731343	0.707463	0.731343	0.707463

Nombre de la fuente: Este trabajo.

Se continúa el análisis con los atributos masa del planeta y radio del planeta en escala logarítmica, obteniendo los resultados de precisión y exactitud de la figura 3-8. Los algoritmos k-NN y árbol de decisiones son los modelos con mayor rendimiento frente al SVM. Comparando estas métricas con la ilustración 3-1 donde los datos se encuentran en escala normal, se observa un incremento mayor a 5 unidades porcentuales en SVM, k-NN y superior al 30% en árbol de decisiones.

Figura 3-8: Métricas con masa y radio del planeta escala logarítmica.

Modelos ML(Datos LOG)	Precisión Masa-Radio	Exactitud Masa-Radio
Máquinas de Vectores	0.952239	0.952239
k-Vecinos más cercanos	0.973134	0.973134
Árbol de decisiones	0.973134	0.973134

Nombre de la fuente: Este trabajo.

Adicionalmente, como se ilustra en 3-9 se toman en escala logarítmica los atributos período orbital y eje semimayor. Con base en la figura 3-3 y 3-10 se puede afirmar que el rendimiento ofrecido por los modelos mejora considerablemente, en particular para el algoritmo árbol de decisiones, cuyo incremento es superior al 30% con un poco más del 15% para SVM y k-NN.

Figura 3-9: Métricas con Periodo orbital y eje semimayor.

Modelos ML(Datos LOG)	Precisión Orbital-Eje	Exactitud Orbital-Eje
Máquinas de Vectores	0.513433	0.513433
k-Vecinos más cercanos	0.510448	0.510448
Árbol de decisiones	0.459701	0.459701

Nombre de la fuente: Este trabajo.

En la tabla de la ilustración 3-10 se pueden observar los terceros modelos estudiados.

Figura 3-10: Métricas comparativas terceros modelos.

Modelos ML(Datos LOG)	Precisión Masa-Radio	Precisión Orbital-Eje	Exactitud Masa-Radio	Exactitud Orbital-Eje
Máquinas de Vectores	0.952239	0.513433	0.952239	0.513433
k-Vecinos más cercanos	0.973134	0.510448	0.973134	0.510448
Árbol de decisiones	0.973134	0.459701	0.973134	0.459701

Nombre de la fuente: Este trabajo.

Para determinar la respuesta de los algoritmos se incluye la tercera característica en escala logarítmica con las métricas obtenidas en la ilustración 3-11, comparando este resultado con lo observado en la figura 3-8, no se presenta mejora significativa en el rendimiento de los modelos predictivos implementados.

Figura 3-11: Métricas con radio, masa y periodo orbital.

Modelos ML (Datos LOG)	Precisión radio-masa-orbital	Exactitud radio-masa-orbital
Máquinas de Vectores	0.955224	0.955224
k-Vecinos más cercanos	0.964179	0.964179
Árbol de decisiones	0.973134	0.973134

Nombre de la fuente: Este trabajo.

Ahora bien, como paso adicional, se incluyen todas las características con el resultado de la ilustración 3-12, con los datos en escala logarítmica. De los modelos predictivos, k-NN presenta en sus métricas un decremento inferior al 2% observando en los demás algoritmos una constante, por lo que el atributo eje semimayor en escala logarítmica sigue sin generar un cambio significativo en el rendimiento de los modelos.

Figura 3-12: Métricas con masa, radio, período orbital y eje semimayor.

Modelos ML (datos LOG)	Precisión Total	Exactitud Total
Máquinas de Vectores	0.955224	0.955224
k-Vecinos más cercanos	0.949254	0.949254
Árbol de decisiones	0.973134	0.973134

Nombre de la fuente: Este trabajo.

De esta manera se sintetizan los resultados obtenidos en la ilustración 3-14. Se concluye que se obtiene un mejor rendimiento o una mejor generalización de los distintos modelos predictivos al implementar los datos en escala logarítmica, se debe resaltar que el atributo eje semimayor no genera un impacto positivo en los algoritmos.

Figura 3-14: Métricas comparativas cuartos modelos.

Modelos ML (datos LOG)	Precisión Parcial	Precisión Total	Exactitud Parcial	Exactitud Total
Máquinas de Vectores	0.955224	0.955224	0.955224	0.955224
k-Vecinos más cercanos	0.964179	0.949254	0.964179	0.949254
Árbol de decisiones	0.973134	0.973134	0.973134	0.973134

Nombre de la fuente: Este trabajo.

Capítulo 4

4.1 Conclusiones

Es importante observar el trabajo de Rincón, de cuya investigación se obtienen al implementar los algoritmos de aprendizaje supervisado k-NN y árbol de decisiones las métricas de precisión y exactitud de 0.93465-0.93465 y 0.948328-0.948328 respectivamente, con una base de datos de 3286 exoplanetas. (Rincón, 2021). Al realizar el modelado de las mismas técnicas de aprendizaje supervisado como se evidencia en el presente trabajo con una base de datos de 1672 instancias, se obtuvo cómo métricas de evaluación valores de 0.964179-0.964179 y 0.973134-0.973134, con una variación estimada de 3 unidades porcentuales con respecto a la primera investigación. Lo expresado indica que, para este tipo de modelos predictivos, una base de datos con un número de instancias superior a un millar y medio se puede considerar como un valor prudente.

Un paso importante, en la construcción de modelos de aprendizaje supervisado, corresponde al pre-procesamiento de los datos, donde las características escogidas para los distintos algoritmos, son desplegadas en gráficas de dispersión, en busca de atributos con alta correlación y agrupamiento de datos. En el presente trabajo, se observó un fenómeno que llamó la atención en un inicio con la gráfica de dispersión de las características período orbital y eje semimayor al desplegar una buena correlación con solapamiento de datos. Como resultado de lo anterior se obtienen idénticas métricas de evaluación al implementar los algoritmos con los atributos masa del planeta, radio del planeta, período orbital y los mismos modelos con la adición del eje semimayor.

Según la literatura, se tiene que las técnicas de aprendizaje supervisado SVM y el algoritmo árbol de decisiones son considerados modelos eficaces en tareas de clasificación y regresión (Bedell, 2018) (Serra, 2020). Al implementar en este trabajo los mencionados clasificadores, se obtienen métricas de 0.955224 y 0.973134 corroborando lo afirmando por los autores.

4.2 Recomendaciones

Las redes neuronales se han constituido en un tema de interés y actualidad, no por ser una idea nueva, es más, las publicaciones con los primeros conceptos datan de los años 40 y 50. El poco reconocimiento logrado en sus inicios, se debía a la falta de disposición tecnológica en recursos computacionales para el entrenamiento y ejecución de estas redes con buenos resultados. (Xalaca, 2016). Este recurso bien entrenado, puede realizar el trabajo de los modelos predictivos implementados en esta investigación, con el potencial de ofrecer resultados más precisos, pero lentos, lo que indudablemente ofrece una línea de investigación en futuros trabajos. (Health Big Data, 2020).

Como queda claro a lo largo del presente trabajo, la gráfica de dispersión de los atributos período orbital y eje semimayor, demuestran una alta correlación con solapamiento de datos. Como investigación adicional se propone sustituir la característica variable eje semimayor por una que despliegue agrupamiento de clases, lo que podría mejorar las métricas de evaluación “precisión y exactitud” logradas en este estudio.

Resulta interesante continuar este estudio con técnicas de aprendizaje no supervisado como k-means, a través de los cuales es posible descubrir los grupos ocultos en los datos (clases de exoplanetas) al determinar los centroides de las gráficas elaboradas, que para el presente caso debe corresponder a las cuatro etiquetas de clasificación (Gigante gaseoso, super tierra, terrestre y tipo Neptuno).

Bibliografía

Análisis y visualización de datos usando Python. (2020). Obtenido de https://datacarpentry.org/python-ecology-lesson-es/02-starting-with-data/

Aprende con Alf. (4 de octubre de 2020). Obtenido de https://aprendeconalf.es/docencia/python/manual/matplotlib/

Aprende Machine Learning . (11 de septiembre de 2017). Obtenido de https://www.aprendemachinelearning.com/7-pasos-machine-learning-construir-maquina/

Aprende machine learning . (2018). Obtenido de https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

aprendeIA. (2021). *K Vecinos mas Cercanos – Práctica con Python.* Obtenido de https://aprendeia.com/k-vecinos-mas-cercanos-programa-con-python-machine-learning/

Astropedia. (2021). Obtenido de https://astronomia.fandom.com/wiki/Uni%C3%B3n_Astron%C3%B3mica_Internacional#:~:text=La%20Uni%C3%B3n%20Astron%C3%B3mica%20Internacional%20(UAI,as%C3%AD%20como%20los%20est%C3%A1ndares%20en

Aura Quantic. (2021). Obtenido de https://www.auraquantic.com/es/que-es-la-inteligencia-artificial/

Aws. (2021). Obtenido de https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/analyzing-your-data.html

B12. (3 de febrero de 2020). Obtenido de https://agenciab12.com/noticia/que-es-modelo-predictivo-como-aplica-negocio

Ball. (2020). Obtenido de https://www.ball.com/aerospace/programs/k2-kepler

Barrios, J. (26 de julio de 2019). *La matriz de confusión y sus métricas*. Obtenido de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

BBVA. (8 de noviembre de 2019). *'Machine learning': ¿qué es y cómo funciona?* Obtenido de

https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/

Bedell, Z. (7 de Diciembre de 2018). Obtenido de https://medium.com/@zachary.bedell/support-vector-machines-explained-73f4ec363f13

Bit Degree. (2020). Obtenido de División de conjuntos de datos con la función train_test_split de Sklearn: https://www.bitdegree.org/learn/train-test-split

CNES. (21 de enero de 2021). *WIKIPEDIA*. Obtenido de Sobre el CNES: https://cnes.fr/en/web/CNES-en/3773-about-cnes.php

DelftStack. (2020). Obtenido de Compruebe los valores de Nan en python: https://www.delftstack.com/es/howto/python/check-for-nan-values-python/#:~:text=El%20nan%20es%20una%20constante,es%20legal%20%2D%20Not%20a%20Number%20.&text=En%20Python%2C%20tenemos%20la%20funci%C3%B3n,comprobar%20los%20valores%20de%20nan%20.

Dipanjan Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical Machine Learning with Python.* Obtenido de https://library.kre.dp.ua/Books/2-4%20kurs/Програмування%20%2B%20мови%20програмування/Python/practical-machine-learning-python-problem-solvers.pdf

Esero spain. (2021). Obtenido de http://esero.es/practicas-en-abierto/exoplanetas-secundaria/resumen_de_los_principales_mtodos_de_deteccin_de_exoplanetas.html

ESO. (2021). Planetas extrasolares. *Planetas extrasolares*. Obtenido de Nasa: https://www.eso.org/public/archives/presskits/pdf/presskit_0004.pdf

García, J. S. (2018). PLANETAS EXTRASOLARES. *Anuario Astronómico del Observatorio* , 391-407.

Globe. (2018). *Datos de prueba*. Obtenido de https://www.globetesting.com/datos-de-prueba/

González, A. (2021). *CleverData*. Obtenido de https://cleverdata.io/conceptos-basicos-machine-learning/

Health Big Data. (15 de Junio de 2020). *Inteligencia Artificial y Machine Learning para todos*.

IArtificial.net. (2021). Obtenido de https://www.iartificial.net/fases-del-proceso-de-machine-learning/

Iberdrola. (2021). Obtenido de https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

Instituto de ingenieria del conocimiento. (2021). Obtenido de

https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Kilodegree Litle. (2019). Obtenido de KELT: https://keltsurvey.org/about

Los exoplanetas. (2021). *Método transito*. Obtenido de http://exoplanetaumce.blogspot.com/p/metodo-del-transito.html

Machine Learning En Español.com. (12 de enero de 2021). *Evaluación de Modelos Predictivos: Ejemplo Python*. Obtenido de https://machinelearningenespanol.com/2021/01/12/evaluacion-de-modelos-predictivos/

Marrugat, A. S. (julio de 2020). *Comparación de algoritmos de clasificación supervisada.* Obtenido de https://upcommons.upc.edu/bitstream/handle/2117/330482/tfm-mueo-alexandre-serra.pdf?sequence=1&isAllowed=y

Martínez, J. (28 de Mayo de 2019). *Máquinas de Vectores de Soport(SVM)*. Obtenido de https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Metodos numericos . (2021). Obtenido de http://blog.espol.edu.ec/analisisnumerico/recursos/resumen-python/graficas-2d-de-linea/

Microsoft. (2021). Obtenido de https://support.microsoft.com/es-es/office/crear-o-editar-archivos-csv-para-importarlos-a-outlook-4518d70d-8fe9-46ad-94fa-1494247193c7

Minitab. (2020). Obtenido de https://support.minitab.com/es-mx/minitab/19/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/all-statistics-and-graphs/

Montero, M. Á. (2009). *Extraccion de conocimiento en bases de datos Astronónicas.* Sevilla.

Morante, S. (1 de noviembre de 2018). *Precauciones a la hora de normalizar los datos en Data Science*. Obtenido de https://empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar/

Muller, A., & Guido, S. (2017). *Introduction to Machine Learning whit Python: tercera edición.*

Nasa. (30 de octubre de 2018). Obtenido de https://www.nasa.gov/mission_pages/kepler/overview/index.html

NASA. (2021). *5 Formas de encontrar un planeta*. Obtenido de https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#

Nasa. (2021). *Exoplanet Exploration* . Obtenido de https://exoplanets.nasa.gov/discovery/exoplanet-catalog/

NASA. (2021). *TESS de la NASA crea una vista cósmica del cielo del norte.* Obtenido de

https://www.mdscc.nasa.gov/index.php/2020/10/06/tess-de-la-nasa-crea-una-vista-cosmica-del-cielo-del-norte/

ORACLE. (2021). *¿Qué son los big data?* Obtenido de https://www.oracle.com/co/big-data/what-is-big-data/

Pacheco, V. G. (18 de Enero de 2019). *Una Breve Historia del Machine Learning*. Obtenido de Telefnica´Tech: https://empresas.blogthinkbig.com/una-breve-historia-del-machine-learning/

Pandas. (2021). Obtenido de https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html

Platzi. (2018). Obtenido de ¿Qué es una librería y para qué sirve?: https://platzi.com/comunidad/que-es-una-libreria-y-para-que-sirve/

REYES, J. G. (2014). *Ajuste simultaneo a curvas de luz y velocidad radial para sistemas en transito.* Obtenido de Ajuste simultáneo a curvas de luz y velocidad radial para sistemas en tránsito

Rincón, K. J. (2021). *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.* Ibagué.

Roman, V. (18 de febrero de 2019). *Medium*. Obtenido de Ciencia y Datos: https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48

Ruiz, N. L.-C. (2017). *Indagación, Exoplanetas y Competencia Científica. Los estudios de Caso como ABP para las Ciencias.* Obtenido de https://raco.cat/index.php/ECT/article/view/328894

Sanchez, F. J. (2021). *Fundamento escalas logaritmicas.* Obtenido de https://hidrologia.usal.es/Complementos/papeles_log/fundamento_log.pdf

Sánchez, V. S. (2019). *Detección de exoplanetas en sistemas binarios.*

Santos, P. R. (28 de abril de 2020). *telefonica* . Obtenido de https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/

SEA. (2021). *Estrella Enana*. Obtenido de https://www.sea-astronomia.es/glosario/estrella-enana

Seaborn. (2021). Obtenido de https://seaborn.pydata.org/

Serra, A. (2020). *Comparacion de algoritmos de clasificacion supervisada* .

Seti. (2021). *NASA's Kepler & K2 Missions*. Obtenido de https://www.seti.org/event/nasas-kepler-k2-missions

Singh, N. (2 de septiembre de 2020). *Métricas de evaluación de modelos en el aprendizaje automático.* Obtenido de https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

SINNETIC Powering Business Through Science. (2009). Obtenido de https://www.sinnetic.com/analitica-de-datos-y-bigdata-para-empresas?utm_source=google.com&utm_medium=organic

Suárez, H. R. (7 de septiembre de 2015). *incibe-cert*. Obtenido de https://www.incibe-cert.es/blog/correlacion-herramientas-analisis-datos

UAI. (10 de octubre de 2019). Obtenido de https://www.iau.org/public/themes/naming_exoplanets/spanish/

UNED. (05 de noviembre de 2019). *LA ASTRONOMÍA: UNA CIENCIA BIG DATA*. Obtenido de https://www.masterbigdataonline.com/index.php/en-el-blog/192-la-astronomia-una-ciencia-big-data

UNIR. (2020). Obtenido de https://www.unir.net/marketing-comunicacion/revista/graficos-estadisticos/

Universidad de Alcalá. (2021). *SCIKIT-LEARN, HERRAMIENTA BÁSICA PARA EL DATA SCIENCE EN PYTHON*. Obtenido de https://www.master-data-scientist.com/scikit-learn-data-science/

Universidad de Arizona. (2021). *UKIRT*. Obtenido de https://www.as.arizona.edu/ukirt

Universidad de Cambridge. (2021). *La cámara de infrarrojos de campo amplio para UKIRT*. Obtenido de http://casu.ast.cam.ac.uk/surveys-projects/wfcam

Wikipedia. (29 de septiembre de 2021). Obtenido de https://es.wikipedia.org/wiki/NASA#:~:text=La%20Administraci%C3%B3n%20Nacional%20de%20Aeron%C3%A1utica,la%20investigaci%C3%B3n%20aeron%C3%A1utica%20y%20aeroespacial.

WIKIPEDIA. (25 de Septiembre de 2021). Obtenido de https://es.wikipedia.org/wiki/Planeta_extrasolar#:~:text=Un%20planeta%20extrasolar%20o%20exoplaneta,cient%C3%ADfica%20en%20el%20siglo%20XX.

WIKIPEDIA. (2021). *Telescopio infrarrojo del Reino Unido*. Obtenido de

https://es.wikinew.wiki/wiki/United_Kingdom_Infrared_Telescope

Xalaca. (21 de Enero de 2016). *Las redes neuronales: qué son y por qué están volviendo.*