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.

Kimberly Johana Rincón Valencia

Universidad Antonio Nariño

Facultad de Ingeniería Mecánica, Electrónica y Biomédica

Ibagué, Colombia

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

Kimberly Johana Rincón Valencia

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

INGENIERO ELECTRONICO

Director (a):

Doctor Alexander Moreno Briceño

Línea de Investigación:

Electrónica Digital

Universidad Antonio Nariño

Facultad de Ingeniería Mecánica, Electrónica y Biomédica

Ibagué, Colombia

2021

Dedicatoria

A mi familia

Quienes con su amor, paciencia y esfuerzo me han guiado y acompañado incondicionalmente a lo largo de este camino y me han impulsado a cumplir hoy una meta más. A ustedes, las personas que nunca me abandonaron y siempre han estado para mí en los momentos más difíciles de mi vida, gracias por inculcar en mí el ejemplo de esfuerzo y valentía, ustedes son el pilar más importante de mi vida.

Agradecimientos

Constantemente poseemos sueños, metas que queremos terminar y que nos ayudan a robustecer nuestra vida personal y profesional, todos dichos fines constantemente necesitan sacrificios y compromisos o no lograríamos la satisfacción de conseguir este gran fin que nos complete de alegría y felicidad. Quiero agradecer a mi familia por el apoyo incondicional y estar en los momentos más difíciles de este gran proyecto de vida. Por la enseñanza recibida y los valores heredados nos ayudan a ser cada vez mejores como personas y humanos, de ahí parte todos nuestros propios sueños e ilusiones, ya que enfocamos nuestros propios fines, somos exigentes y deseamos mejorar día a día y gracias al apoyo de ellos logramos sortear nuestros propios senderos dirigidos a un solo objetivo ser ingeniera electrónica.

Agradezco a la Universidad Antonio Nariño por haberme aceptado ser parte de ella y abierto las puertas de su seno científico para poder estudiar mi carrera, así como también a los diferentes docentes que brindaron sus conocimientos y su apoyo para seguir adelante día a día.
Agradezco también a mis asesores de proyecto de grado por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento científico.

Para finalizar, también agradezco a mis amigos y a todos los que fueros mis compañeros de clase durante los niveles de universidad ya que gracias al compañerismo, amistad y apoyo moral han aportado en un alto porcentaje a mis ganas de seguir adelante en mi carrera profesional.

Gracias.

Resumen

Las técnicas y algoritmos de ML se han vuelto muy importantes en el campo de la astronomía y en muchos otros campos, debido al enorme volumen de imágenes o datos recolectados empleando diferentes medios, derivados del estudio realizado en cada área del conocimiento y que genera la necesidad de trabajar con grupos de datos extensos (Big Data) y complejos para procesarlos, analizarlos y posteriormente realizar observaciones que permiten extraer nuevos conocimientos para predecir comportamientos a futuro de dicha información.

Los datos astronómicos sobre exoplanetas recolectados a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt, se encuentran contenidos en los archivos de exoplanetas de la NASA. Dicha base de datos requiere ser tratada para organizar y homogenizar la información contenida en ella, además de elegir los parámetros o atributos más relevantes que se emplearán en el análisis. Dicho análisis se realizará sobre la información procesada aplicando las técnicas de aprendizaje supervisado y no supervisado de ML para generar un análisis eficiente de la información que culmine con un modelado pertinente que ayude a encontrar patrones y reglas que permitan generar nuevas observaciones para clasificar cada exoplaneta registrado con predicciones a futuro del comportamiento que presenta la información.

Los resultados obtenidos implementando las técnicas y algoritmos de aprendizaje supervisado y no supervisado, permiten estimar a partir de diversos atributos una etiquetación o predicción del tipo de planeta en el que clasifica el dato con una exactitud y precisión bastante alta al incorporar una base de datos extensa o Big Data.

Palabras clave: Algoritmos, Machine Learning, Exoplanetas, Astronomía, Big Data, Nasa.

Abstract

ML techniques and algorithms have become very important in the field of astronomy and in many other fields, due to the enormous volume of images or data collected using different means, derived from the study carried out in each area of knowledge and that generates the need from working with large and complex data groups (Big Data) to process, analyze and later make observations that allow extracting new knowledge to predict future behavior of said information.

The astronomical data on exoplanets collected through various terrestrial and space telescopes and different missions carried out, including: Tess, Kepler, K2, KELT and Ukirt, are contained in the NASA exoplanet archives. This database needs to be processed to organize and standardize the information contained in it, in addition to choosing the most relevant parameters or attributes to be used in the analysis. Said analysis will be executed on the processed information applying ML supervised and unsupervised learning techniques to generate an efficient analysis of the information that culminates in relevant modeling that helps find patterns and rules that generate new observations to classify each exoplanet registered with future predictions of the behavior presented by the information.

The results obtained by implementing the techniques and algorithms of supervised and unsupervised learning, allow estimating from various attributes a labeling or prediction of the type of planet in which the data is classified with a fairly high accuracy and precision by incorporating an extensive database or Big Data.

Keywords: Algorithms, Machine Learning, Exoplanets, Astronomy, Big Data, NASA.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

En la actualidad la necesidad de analizar y procesar gran cantidad de datos, ha impulsado al desarrollo de la ciencia basada en datos como una herramienta para el análisis, procesamiento y estudio detallado de Big data en modelados comunes (BSA data study, 2015). Debido a esto las técnicas y algoritmos de ML se han vuelto tan populares en el campo de la astronomía ya que todos los grandes centros astronómicos utilizan una gran cantidad de información o datos complejos para realizar su investigación de una manera óptima. Dado el aumento considerable en los datos recolectados cada año por observatorios astronómicos en el mundo, estas técnicas y algoritmos actualmente se han vuelto indispensables para poder tratar esa enorme cantidad de información, siendo empleadas en múltiples tareas y en diversas áreas que presentan esta misma tendencia. (CASTILLO, 2018)

Este documento se centra en los algoritmos de aprendizaje automático supervisado y no supervisado empleados en astronomía para el análisis de conjuntos de datos sobre exoplanetas recolectados a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt. (NASA, 2021).

Al analizar esta gran base de datos de la NASA sobre todos los exoplanetas observados con las sondas mencionadas anteriormente, se busca realizar un modelo predictivo empleando las técnicas de ML que permitan clasificar cada uno de los exoplanetas confirmados en los diferentes tipos de exoplanetas hallados hasta el momento, los cuales son: súper tierra, gigante gaseoso, tipo Neptuno y terrestre. (ExoPlanet Exploration, 2021).

La exactitud del modelado recae principalmente en la cantidad y calidad de los datos, análisis y procesamiento de los mismos antes de implementar las técnicas de aprendizaje supervisado y no supervisado. El análisis exploratorio de los datos es una de las fases de ML que supone un mayor esfuerzo y que requiere bastante atención. Siendo la fase en la

que se busca tratar los datos contenidos en la base de datos inicial o de partida para analizarlos, entender su naturaleza y tomar decisiones a partir de dichos procesos y así saber con certeza si se ajustan o no a la naturaleza del problema o hipótesis. También se debe considerar al implementar los diversos algoritmos de ML, ingresar como parámetros de entrada los atributos que aporten mayor información relevante que permitan realizar una correcta estimación o predicción de las etiquetas asignadas a los datos de entrada. Normalizar los datos, organizando la base de datos en una escala similar puede ser muy útil para trabajar mejor con la información recolectada.

Otro paso indispensable en el proceso de análisis y preprocesamiento de la información es el trabajo de ingeniería que permite identificar los atributos que discriminen o aporten más información al proceso de generación del modelo de clasificación, permitiendo así un mayor porcentaje de verdaderos positivos a la hora de clasificar los diversos exoplanetas.

Para la siguiente fase, al tener la base de datos correctamente procesada se debe optar por un algoritmo de aprendizaje supervisado o un aprendizaje no supervisado para implementar. El algoritmo escogido aprenderá automáticamente al obtener los resultados adecuados con los datos históricos que se dispongan o que se le entreguen inicialmente, con un error que se debe trabajar en la siguiente fase para identificar si la predicción se está realizando correctamente y con un margen de error mínimo o si se requiere iterar y realizar un nuevo modelamiento o implementar otro algoritmo que aumente la eficiencia del modelo predictivo obtenido (Roman, 2019).

Evaluar la medida del uso de los recursos computacionales requeridos por la ejecución de un algoritmo de diversos algoritmos de machine learnig a partir del modelado inicial será de gran importancia para definir cuál es el más eficiente para este problema en particular.

Se emplearan diversas métricas a lo largo de la siguiente fase. La fase de análisis de la exactitud del modelado generado para determinar el margen de error obtenido en la generalización que realiza el algoritmo. El tipo de métrica y la metodología empleada para implementarla, dependerá de la técnica y del algoritmo de ML empleado en cada caso.

En la fase del análisis del error obtenido, se busca mejorar los resultados obtenidos. En esta fase se logra definir si el modelo generado es capaz de realizar correctamente

generalizaciones y predicciones o si por el contrario se requiere un modelo más complejo, si la base de datos inicial tiene la cantidad suficiente de información o si las características definidas son correctas, entre otros aspectos que permitan generar un modelo con buenos resultados predictivos al implementar datos nuevos o de prueba. En esta fase se define si se requiere realizar una iteración sobre las fases anteriores para mejorar los datos o características analizadas, si el algoritmo implementado es el que mejor o si se ajusta al problema y conseguir reducir el error de predicción (Roman, 2019).

Por último se busca analizar los algoritmos implementados y determinar cuál de todos es mejor al estimar patrones no identificados previamente. Para realizar este análisis, se emplearan gráficas y tablas que respalden las conclusiones generadas a partir de los resultados obtenidos en cada modelado generado.

En el capítulo 1 se estudiaran las diferentes técnicas de ML haciendo énfasis en algunos algoritmos de aprendizaje supervisado y no supervisado, se definió lo que es un exoplaneta y los diferentes métodos de observación empleadas para la recopilación de la información. En el capítulo 2 se presenta la reproducción de datos y la limpieza, filtrado y preprocesamiento de los mismos. En el capítulo 3 se estudia específicamente las técnicas de aprendizaje supervisado y no supervisado a implementar tales como: Árbol de decisión, k vecinos más cercanos, entre otros. En el capítulo 4 se encuentran los resultados y análisis de los mismos. Por último, en el capítulo 5 están las conclusiones y recomendaciones referentes al proyecto.

Objetivos

Objetivo general

- Implementar técnicas y algoritmos de aprendizaje supervisado y no supervisado de ML en el análisis de datos astronómicos sobre exoplanetas obtenidos por misiones espaciales de la NASA por medio de las sondas Tess, Kepler, K2, KELT y Ukirt.

Objetivos específicos

- Implementar los algoritmos de aprendizaje supervisado, tales como clasificación y regresión, usando datos de archivos de la NASA obtenidos las sondas Tess, Kepler, K2, KELT y Ukirt.
- Implementar los algoritmos de aprendizaje no supervisado, tales como agrupamiento y reducción dimensional, usando datos de archivos de la NASA obtenidos por las sondas Tess, Kepler, K2, KELT y Ukirt.
- Analizar los algoritmos implementados y determinar cuál de todos es mejor al estimar patrones no identificados previamente.

Definiciones

- **Machine learning:** Es el campo de estudio que le da a las computadoras la capacidad de aprender sin ser programadas explícitamente (Dera & Bhavsar, 2017).
- **Aprendizaje supervisado:** Técnica de ML que emplea etiquetas para realizar un entrenamiento del modelo y realizar predicciones del comportamiento esperado a partir de los datos iniciales etiquetados o datos de entrenamiento.
- **Aprendizaje no supervisado:** Técnica de ML que no emplea etiquetas o conocimiento a priori para realizar las predicciones del comportamiento esperado a partir del modelo creado.
- **Sondas:** Son dispositivos robóticos enviados al espacio para realizar estudios o tomar datos determinados de cuerpos del sistema solar. (Exoplanet Exploration, 2021)
- **Exoplanetas:** Son aquellos planetas que no hacen parte de nuestro sistema solar y orbitan alrededor de otras estrellas. (Exoplanet Exploration, 2021)
- **Gigante gaseoso:** Es un gran planeta compuesto principalmente de helio y / o hidrógeno. Estos planetas, como Júpiter y Saturno en nuestro sistema solar, no tienen superficies duras y en cambio tienen gases arremolinados sobre un núcleo sólido. (Exoplanet Exploration, 2021)
- **Tipo Neptuno:** Los exoplanetas neptunianos son similares en tamaño a Neptuno o Urano en nuestro sistema solar. Los planetas neptunianos suelen tener atmósferas dominadas por hidrógeno y helio con núcleos o rocas y metales más pesados. (Exoplanet Exploration, 2021)
- **Súper tierra:** Las súper-Tierras, una clase de planetas diferente a cualquiera de nuestro sistema solar, son más masivas que la Tierra pero más livianas que gigantes de hielo como Neptuno y Urano, y pueden estar hechas de gas, roca o

una combinación de ambos. Tienen entre el doble del tamaño de la Tierra y hasta 10 veces su masa. (Exoplanet Exploration, 2021)

- **Terrestre:** En nuestro sistema solar, la Tierra, Marte, Mercurio y Venus son planetas terrestres o rocosos. Para los planetas fuera de nuestro sistema solar, aquellos entre la mitad del tamaño de la Tierra y el doble de su radio se consideran terrestres y otros pueden ser incluso más pequeños. (Exoplanet Exploration, 2021).

1. Capítulo 1

1.1 Marco teórico

1.2 Machine Learning (ML)

Vivimos en una época en donde existe en abundancia una gran cantidad de datos estructurados y no estructurados en muchas áreas del conocimiento, una abundancia de datos que hace que sea imposible analizarlos, deducir patrones y plantear modelos a partir de ellos, manualmente. De estos datos se puede obtener valiosa información que permite realizar predicciones y basados en estas predicciones, tomar decisiones.

Estas técnicas permiten buscar y encontrar información de una forma más precisa dentro de un gran conjunto de datos o Big Data, deducir reglas y encontrar patrones que permitan construir modelos, algo que sería demasiado complejo para que el ser humano lo realice manualmente (Raschka & Mirjalili, 2019).

Hoy en día, las aplicaciones de las técnicas y algoritmos de aprendizaje automático son amplias y variadas en la industria y en la academia. Un área de conocimiento en particular que ha tenido un importante crecimiento en la cantidad de datos y en su complejidad es la astronomía.

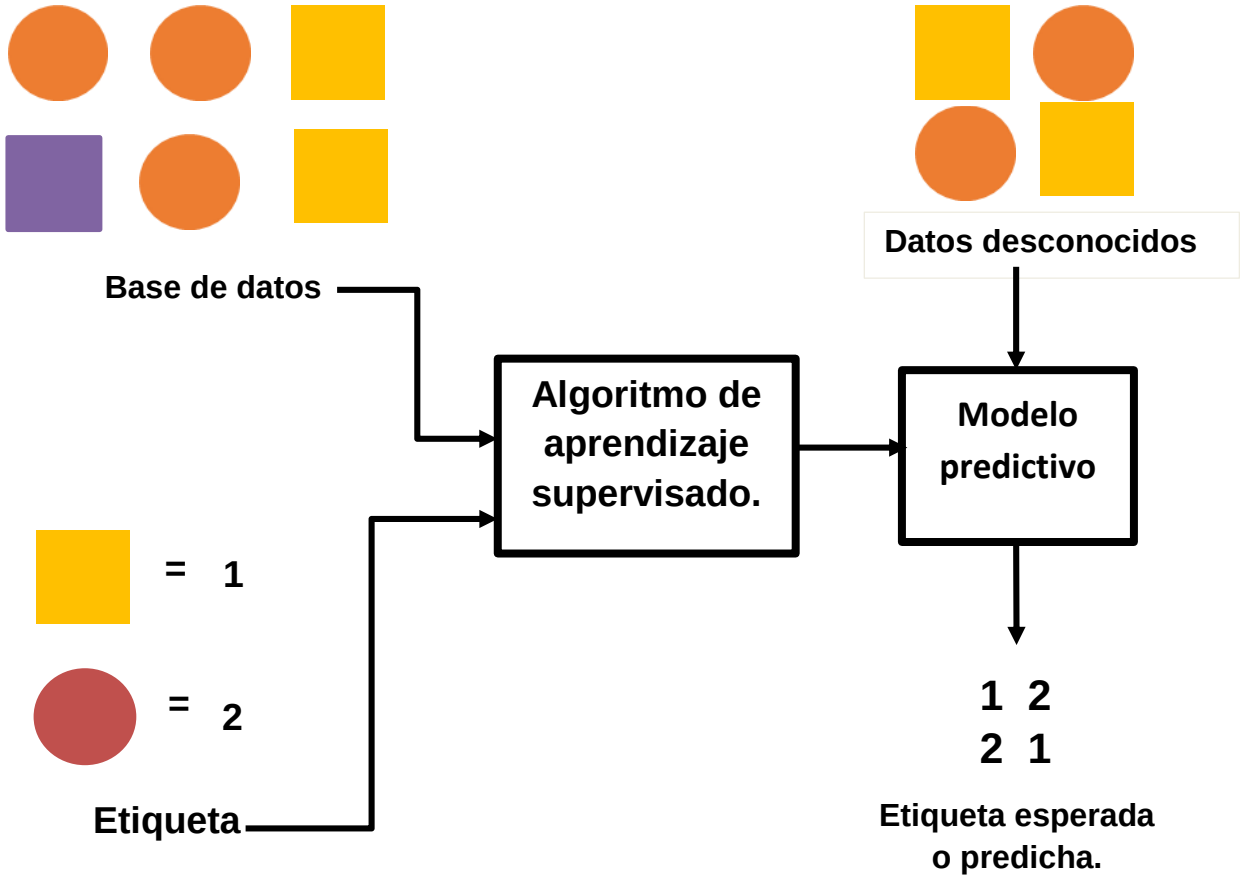
Estos algoritmos se clasifican según las necesidades del problema que se requiere abarcar, el ambiente de trabajo en el que se van a desenvolver dichos algoritmos, además de los múltiples factores que llegaran a afectar de alguna manera la toma de decisiones que se realice en el proceso. Este proyecto se centrará en dos tipos de algoritmos de aprendizaje: Supervisado y no Supervisado.

1.2.1 Aprendizaje supervisado

En el aprendizaje supervisado se dispone de una variable objetivo observada en los datos, empleada como una etiqueta que permite realizar un entrenamiento con los algoritmos implementados y de esta manera realizar la clasificación de la información recolectada para el modelo predictivo de aprendizaje automático. (Machine Learning,una pieza clave en la transformación de los modelos de negocio, 2018).

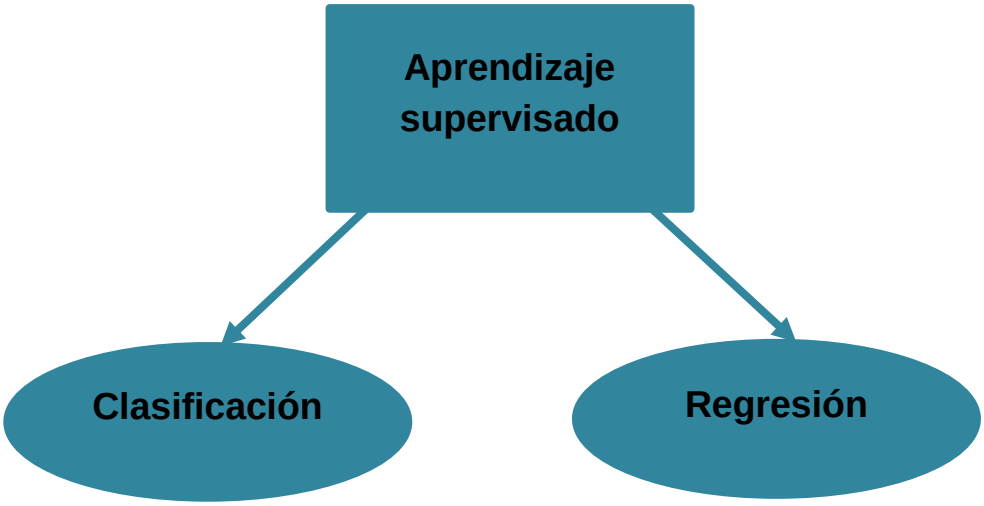
En la figura 1-1 se observa el esquema general de aprendizaje supervisado donde en la entrada se aprecia el ingreso de las etiquetas “1,2” para realizar el entrenamiento y clasificar la nueva información ingresada con el algoritmo implementado en cada caso.

Figura 1-1: Esquema general de aprendizaje supervisado.



En este tipo de aprendizaje hay dos algoritmos o entrenamientos: El algoritmo de clasificación y algoritmo de regresión, como se observa en la figura 1-2.

Figura 1-2: Algoritmos o entrenamientos de aprendizaje supervisado.



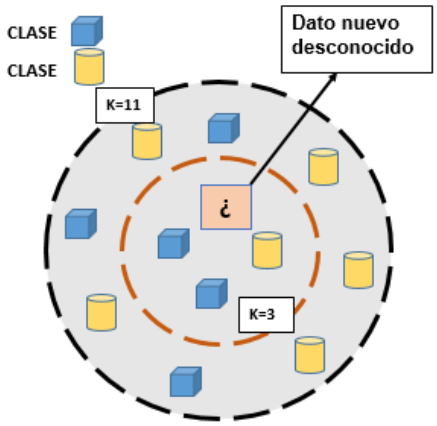
El algoritmo de clasificación encuentra patrones en los datos que se le entregan inicialmente y los clasifica en grupos según las etiquetas entregadas para posteriormente realizar la comparación de nuevos datos ingresados al algoritmo y ubicarlos en uno de los grupos existentes (etiquetas). Por otro lado el algoritmo de regresión entrega como resultado un número, a diferencia del algoritmo de clasificación no lo ubica en un grupo, sino que devuelve un valor específico. (Sandoval, 2018).

Es importante tener en cuenta que las técnicas de aprendizaje supervisado solo funcionan si existe un conjunto de datos históricos que contiene valores reales para el resultado que se intenta predecir con el modelado generado.

Algoritmo K Nearest Neighbour.

El algoritmo de los k vecinos más cercanos (k-NN) es un algoritmo de clasificación de ML supervisado que está basado en criterios de vecindad. Este algoritmo se encarga de clasificar los nuevos ejemplos o datos ingresados con la misma clase o etiqueta que tengan la mayor cantidad de vecinos más parecidos del conjunto de entrenamiento. Este algoritmo primero revisa las similitudes con los datos que ya se clasificaron o etiquetaron anteriormente y por último agrega la misma etiqueta a los datos nuevos ingresados que cumplen con dichas características o atributos (similitudes). Este algoritmo ingresa inicialmente una condición que debe cumplir entre los datos que se ingresen para medir las similitudes. (Caparrini, 2020).

Figura 1-3: Algoritmo K Nearest Neighbour.



En la figura 1-3 se puede observar como realiza la clasificación este algoritmo. El nuevo ejemplo de clasificación o elemento de prueba recibe la misma etiqueta que los k elementos más cercanos del conjunto de entrenamiento. Donde k es el parámetro que indica el número de vecinos que el algoritmo debe considerar a la hora de realizar la clasificación.

En este caso si k = 11, se comprueba las etiquetas de las siete elementos más cercanos y se asigna la etiqueta más común ( ).

Árboles de Decisiones

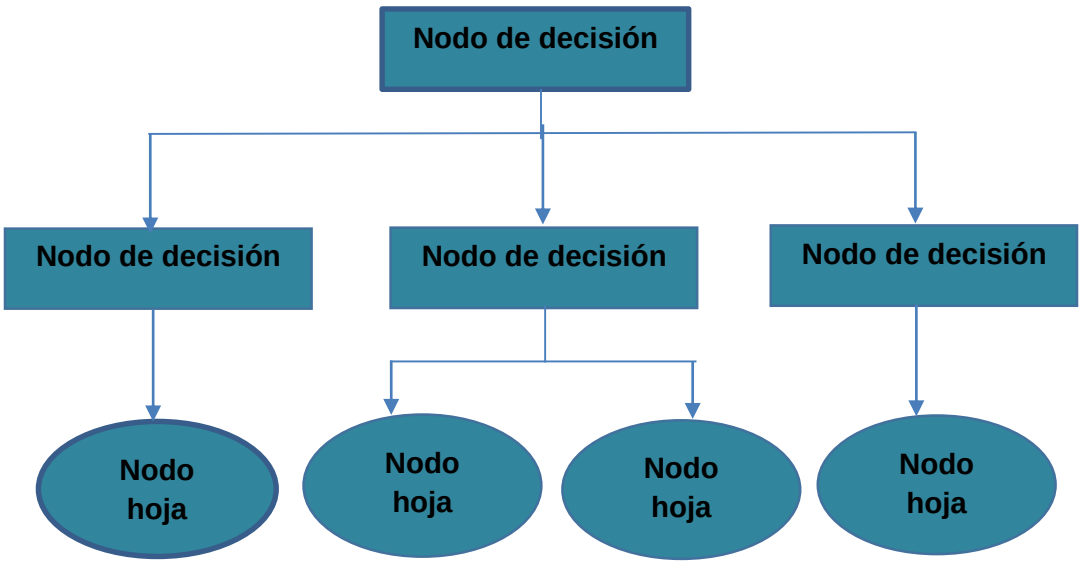
Los árboles de decisiones son algoritmos y modelos más sencillos y fáciles de interpretar aunque con una exactitud más limitada que los demás modelos debido a la partición ortogonal del espacio que realiza y al sobre entrenamiento que este tipo de algoritmos suele presentar, aunque su precisión también depende en gran medida de los datos de entrada que se le otorgan al algoritmo y el problema que se esté modelando. (Management Solutions, 2018).

El nombre que se le otorga a este algoritmo hace referencia a su estructura de árbol donde podemos encontrar una raíz y unos nodos, también se puede observar las ramas y las hojas. Este modelo empieza por el nodo raíz y se extiende hacia abajo en dos o más ramas según se requiera de izquierda a derecha. El nodo donde culmina la estructura es el nodo hoja. (RUIZ, 2018)

Cada nodo intermedio del árbol se considera como un atributo o una característica que se posee de los datos trabajados o datos de entrada, el nodo raíz es el atributo identificado como el más relevante, y las hojas son las correspondientes etiquetas que se tengan. Se realiza un agrupamiento de los datos o etiquetado de los mismos según los atributos o características que comparten entre ellos. (RUIZ, 2018)

La estructura del esquema básico de un árbol de decisiones se puede observar en la figura 1-4.

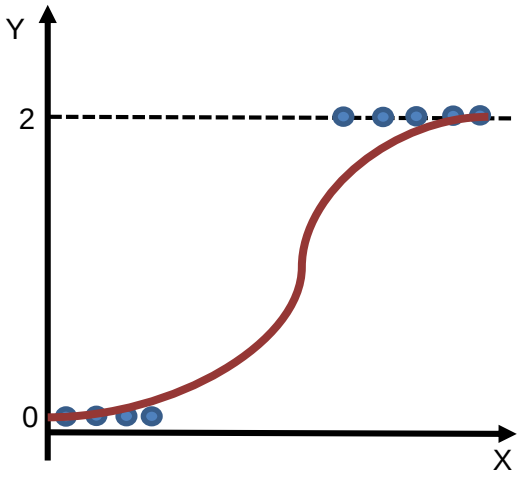
Figura 1-4: Estructura general de un Árbol de decisión.



Algoritmo de Regresión Logística

El algoritmo de Regresión Logística es un modelo para clasificación muy simple de implementar aunque es difícil que converja si los datos que se están tratando no son separables linealmente. En estadística la regresión Logística es un modelo donde la variable dependiente que se defina es categórica y puede tomar valores fijos o un rango de valores posibles. Cuando se habla de variables dependientes se hace referencia a aquellas variables definidas que se desean predecir con el modelo, mientras que las variables independientes son todas las características que se tengan en el conjunto de datos y que se emplean para realizar el entrenamiento del modelo predictivo. Para entender el funcionamiento de este algoritmo es necesario entender el ratio de probabilidad o la probabilidad con la que puede ocurrir un evento. El ratio de puede definir como: p 1−p, donde p representa la probabilidad de que ocurra un evento positivo dentro del modelo y se indica con qué certeza se va a producir un evento. En un modelo predictivo de clasificación con múltiples clases existe la probabilidad de que un dato pertenezca o se pueda clasificarse dentro de las posibles clases o etiquetas disponibles. (RUIZ, 2018).

Figura 1-5: Grafico regresión logística**.**

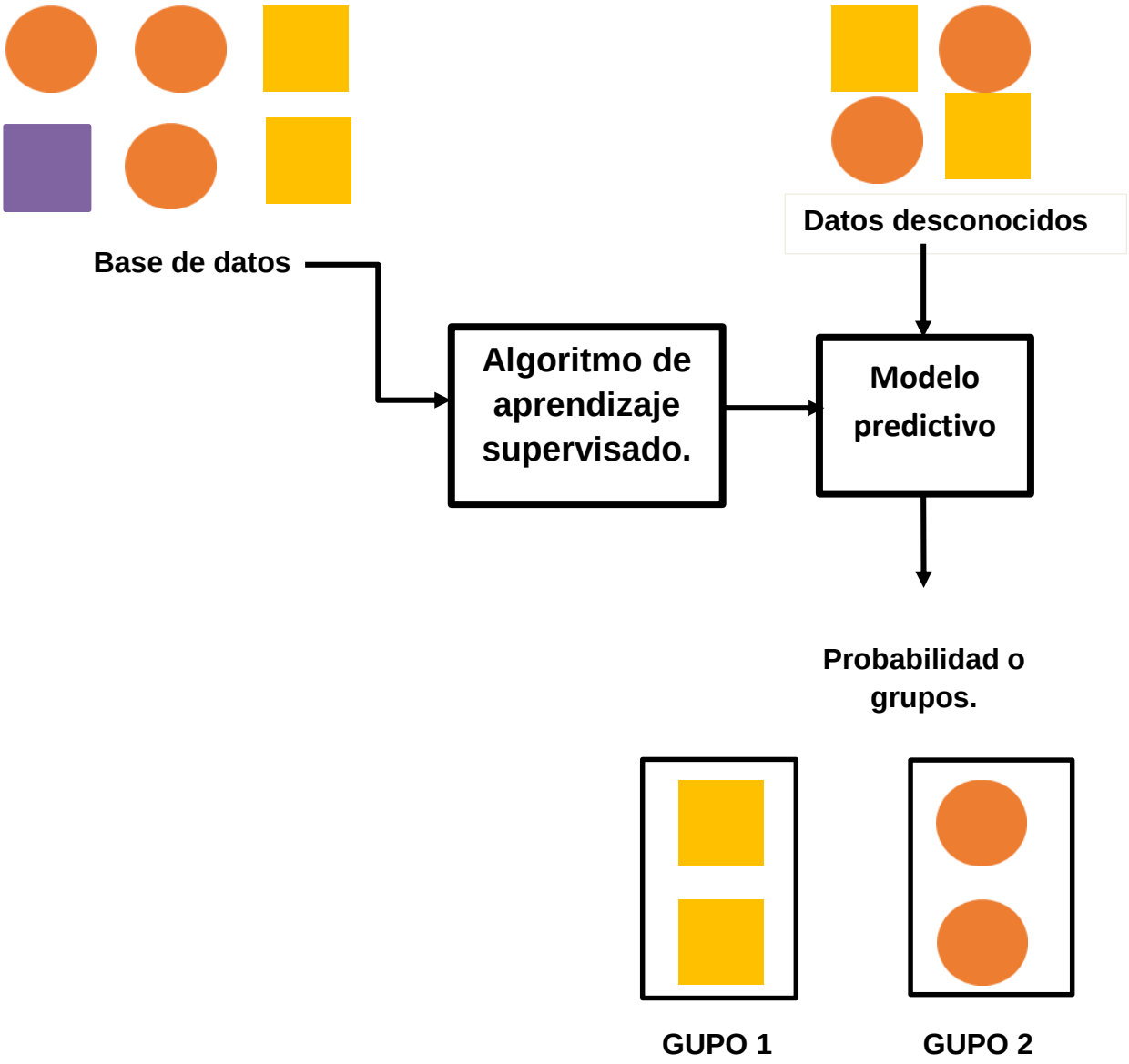


El objetivo de la regresión logística es construir un modelo que prediga la clase de cada observación y la probabilidad de que cada dato pertenezca a una clase. En la figura 1-5 se realiza la predicción de Y que se encuentra dentro del rango de 0 y 2.

1.2.2 Aprendizaje no supervisado

En el aprendizaje no supervisado el objetivo es encontrar patrones o relaciones en los datos recolectados sin implementar un variable objetivo o una etiqueta a la hora de entrenar los algoritmos de ML implementados. (Machine Learning,una pieza clave en la transformación de los modelos de negocio, 2018).

Figura 1-6: Esquema básico de aprendizaje no supervisado**.**

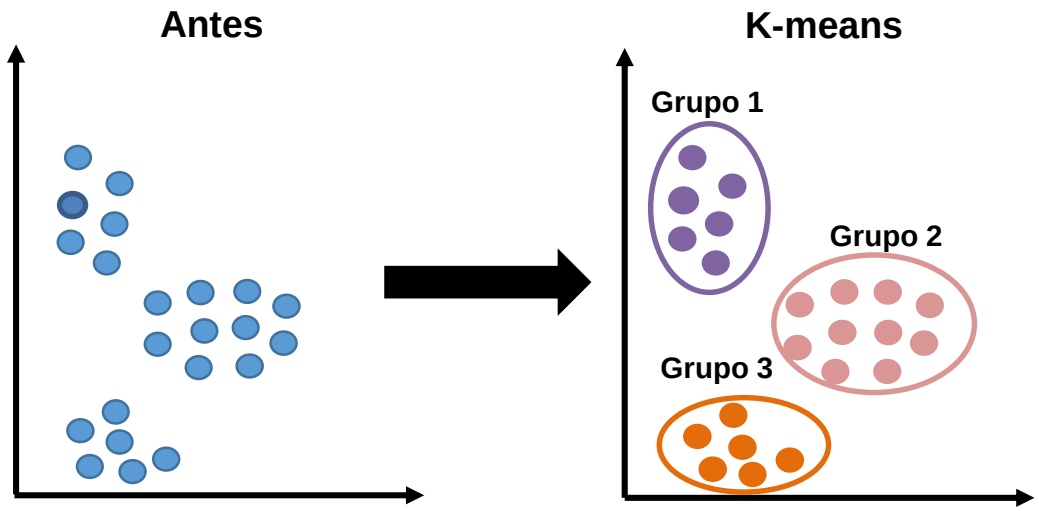


En el esquema básico del aprendizaje no supervisado de la figura 1-6, se observan los datos de entrada al algoritmo junto con los atributos relevantes. También se observa el ingreso de los nuevos datos y sus atributos para generar la predicción con el modelo a partir de la agrupación por clústeres o grupos.

Algoritmo K-means

El algoritmo de las k-medias o k-means de aprendizaje No supervisado, es aplicable en los casos que se cuente con una representación de los datos como elementos en un espacio métrico. k-medias busca encontrar una partición de las muestras en k agrupaciones, de forma que cada ejemplo pertenezca a una de ellas, concretamente a aquella cuyo centroide esté más cerca. El mejor valor de k se debe hallar para que la clasificación separe lo mejor posible la información. Las referencias para el modelo predictivo no se conocen a priori, y depende completamente de los datos de entrada o iniciales. En este caso no se cuenta con un conocimiento a priori que nos indique cómo deben agruparse ninguno de los datos que sea ingresado como en el caso del aprendizaje supervisado que cuenta con las etiquetas correspondientes. (Caparrini, 2020).

Figura 1-7: Clustering con k-means.



En la figura 1-7 se observa la agrupación o los clúster generados después de encontrar una partición de las muestras en k=3 agrupaciones, para este caso en concreto.

Algoritmo Clustering Jerárquico

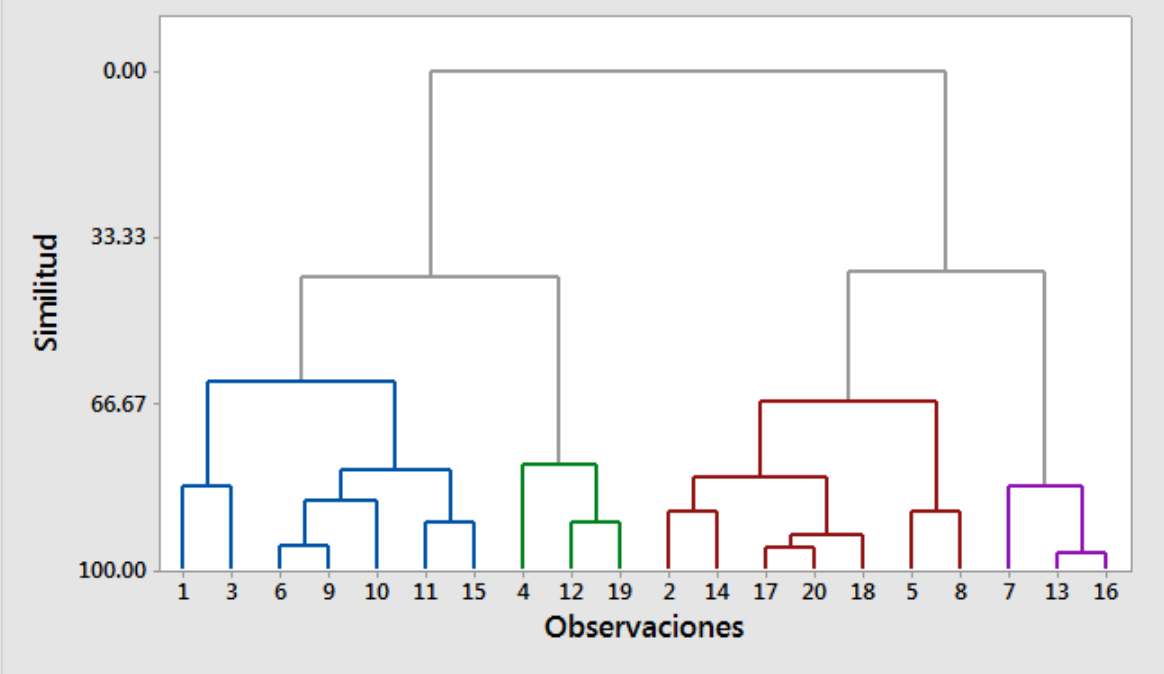
El algoritmo clustering jerárquico de aprendizaje no supervisado, se visualiza empleando un dendograma, donde cada agrupación es representada por una línea horizontal. En el eje x se encuentra la similitud entre los clusters o grupos que se generaron y las observaciones en el eje Y se visualizan como clusters individuales. Un dendograma

permite reconstruir la historia de agrupamientos que resultaron en el clustering representado. (HOJAS, s.f.)

Se utiliza el dendrograma para observar cómo se forman los grupos en cada paso y para evaluar los niveles de similitud (o distancia) de los Clusters que se forman. La decisión acerca de la agrupación final se conoce como cortar el dendrograma. Cortar el dendrograma es trazar una línea a lo largo del mismo para especificar el número de Clusters a formar o las conglomeraciones finales. (Soporte de miniTab18, 2019).

El esquema básico de un dendograma se puede observar en la figura 1-8 mostrada a continuación.

Figura 1-8: Estructucta general de un dendograma.



Nombre de la fuente: Soporte de miniTab18 dendograma. Recuperado de: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/**.**

El método Ward para el clustering jerarquico indica que la perdida de información al realizar el proceso de división o agrupamiento bajo clusters puede medirse a través de la suma total de los cuadrados de las desviaciones entre cada punto (dato o planeta) y la media del grupo que lo contiene (Venegas & Pineda Rios, 2017).

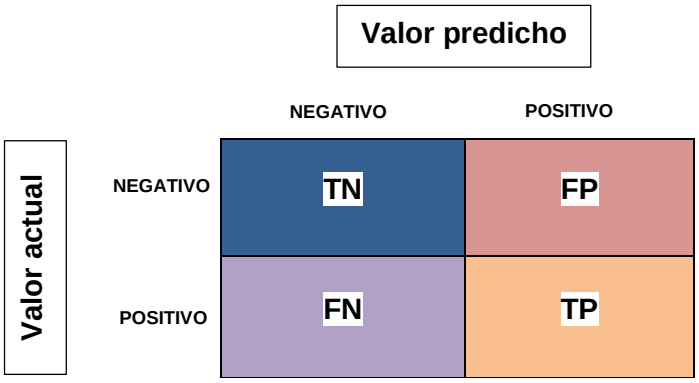
El objetivo de este trabajo es el de implementar los algoritmos mencionados anteriormente de ML en Astronomía para el análisis de datos astronómicos sobre exoplanetas obtenidos por misiones espaciales de la NASA por medio de las sondas Tess, Kepler, K2, KELT y Ukirt, para encontrar patrones y reglas que nos permitan realizar nuevas observaciones.

1.2.3 Métricas de evaluación del modelo.

La preparación, el filtrado y preprocesamiento de la información, además del entrenamiento del modelo generado de aprendizaje automático es un paso clave en el proceso de generar un modelo predictivo eficiente, pero también se debe considerar igualmente importante la medición del rendimiento de dicho modelo generado. Lo eficiente que el modelo sea a la hora de realizar generalizaciones sobre los datos nuevos ingresados al algoritmo, es lo que define si el modelo es adaptable o no adaptable. Por lo mencionado anteriormente, es importante emplear diversas métricas que nos permitan evaluar el rendimiento para lograr conseguir por medio de esta información adquirida, mejorar la exactitud de la predicción general del modelo antes de seguir ingresando nuevos datos al algoritmo. Este paso es demasiado importante a la hora de minimizar el margen de error al ingresar nuevos datos desconocidos a el modelo y de esta forma evitar las malas predicciones que se podrían generar con el despliegue de dichos datos a futuro. (Chauhan, 2020).

La matriz de confusión es una representación matricial de los resultados obtenidos en el modelo predictivo de cualquier prueba binaria y se emplea para lograr calcular y observar el rendimiento de dicho modelo de clasificación con respecto a los datos de prueba que se escogieron inicialmente para ingresar al algoritmo inmediatamente después de realizar el entrenamiento del mismo, como se observa en la figura 1-9 a continuación.

Figura 1-9: Matriz de confusión con 2 etiquetas de clase.



En la figura 1-9 de la matriz de confusión, el grado de la matriz varía según el número de etiquetas que posea el modelo a evaluar y cada espacio de la figura1-9 representa respectivamente (Chauhan, 2020):

- ➔ True Positive (TP): Valor predicho Verdadero y Verdadero en realidad.
- ➔ True Negative (TN): Valor predicho Falso y Falso en realidad.
- ➔ False Positive (FP): Valor predicho verdadero y falso en la realidad.
- ➔ False Negative (FN): Valor predicho falso y verdadero en la realidad.

La exactitud es una métrica de medición que indica la frecuencia con la que es correcto el modelo clasificador y se expresa de la siguiente manera (Chauhan, 2020):

$$\boldsymbol{Exactitud} = \frac{\boldsymbol{Valor\ predicho\ verdadero(TP)+Valor\ perdicho\ falso(TN)}}{\boldsymbol{Total}} \quad (1)$$

Por otro lado, la precisión es la métrica de evaluación que indica el número de predicciones correctas halladas por el modelo predictivo como una proporción de todas las predicciones hechas por el modelo y se expresa de la siguiente manera (Chauhan, 2020):

$$\boldsymbol{Precisión} = \frac{\boldsymbol{Valor\ predicho\ verdadero(TP)}}{\boldsymbol{Valor\ predicho\ verdadero(TP)+valorpredicho\ verdadero\ y\ falso(FP)}} \quad (2)$$

1.3 Exoplanetas

Los exoplanetas son planetas que no se encuentran dentro de nuestro sistema solar, cuyo estudio es de gran importancia en el campo de la astronomía en la búsqueda de planetas

ubicados cerca a otras estrellas con características similares a las de la tierra y que podrían albergar vida. También existen exoplanetas que no se encuentran cerca de una estrella y se hallan flotando libremente en el espacio, conocidos como planetas errantes.

Cuando se realiza la medición de diversos parámetros como el diámetro y la masa de los planetas, se puede conocer a través de esta información la composición de los mismos, que va desde los muy rocosos como la tierra, hasta los muy ricos en gas como Saturno. A pesar de que los exoplanetas están formados por elementos muy similares a los planetas que se encuentran dentro de nuestro sistema solar, sus mezclas pueden ser distintas. (Exoplanet Exploration, 2021)

Los nombres de los exoplanetas son largos y están compuestos por diversas partes que brindan bastante información sobre el mismo y que es importante para la forma en que los científicos catalogan miles de planetas. Los astrónomos diferencian entre las "designaciones" alfanuméricas y los "nombres propios" alfabéticos. Todas las estrellas y exoplanetas tienen designaciones, pero muy pocos tienen nombres propios como por ejemplo: Tierra y Venus. (Exoplanet Exploration, 2021)

Para comenzar, la primera parte del nombre puede ser la sonda o telescopio que lo descubrió y el número es el orden de catalogamiento de la estrella por posición. La tercera parte es la letra minúscula de representación del planeta y en el orden de hallazgo empezando por la b y continuando en orden alfabético, comenzando por el más cercano y finalizando con el más distante con respecto a la estrella que orbita. La estrella que orbita el exoplaneta suele ser la "A" no declarada del sistema, teniendo en cuenta que las estrellas se denotan con mayúscula y los planetas con minúscula. Esta información puede ser útil si el sistema contiene muchas estrellas. Un ejemplo es Kepler-16b, donde "Kepler" es el nombre del telescopio que lo descubrió, 16 es el orden en el que se catalogó la estrella y "b" es el planeta más cercano a la estrella (NASA, 2021).

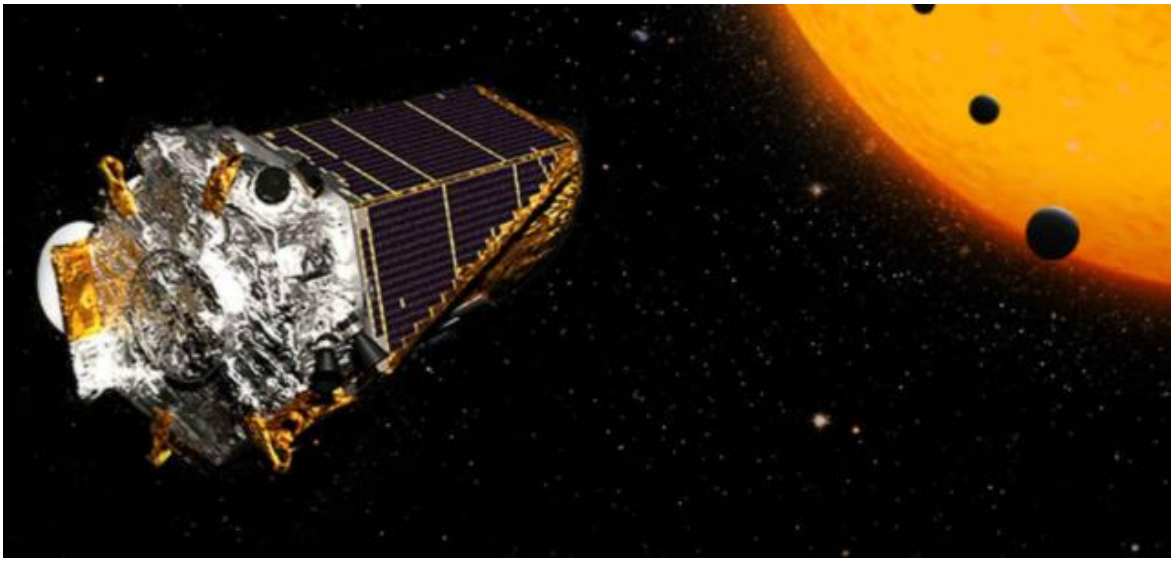
1.3.1 Instrumentos de Observación

A continuación se mencionan algunas de las misiones, telescopios y estudios empleados en la detección de planetas extrasolares.

Kepler

La misión Kepler, fue una misión discovery de la NASA lanzada el 6 de marzo de 2009, fue la pionera en lo que corresponde a misiones espaciales dedicadas a la búsqueda de exoplanetas en nuestras galaxias vecinas. Kepler monitorea continuamente más de 100,000 estrellas similares a nuestro sol para detectar cambios de brillo producidos por los tránsitos planetarios. La misión de la sonda Kepler que se observa en la figura 1-10 renació como la misión K2, que duró cinco años más. (NASA, 2021).

Figura 1-10: Kepler telescopio espacial en busca de exoplanetas.



Nombre de la fuente: Omicrono. **Recuperado de:** https://www.elespanol.com/omicrono/tecnologia/20181031/nasa-retira-satelite-kepler-repasamos-grandes-logros/349716303_0.html

K2

La misión K2 continuó con los descubrimientos y estudios realizados por la nave espacial Kepler en el campo de los exoplanetas y extendió sus observaciones astrofísicas a lo largo de su funcionamiento. La misión K2 finalizó sus observaciones el 30 de octubre de 2018, tras quedar sin combustible la nave. (NASA, 2021).

UKIRT

La cámara de campo amplio (WFCAM) del telescopio infrarrojo del Reino Unido (UKIRT) que se observa en la figura 1.11, se encuentra ubicada en el Observatorio de Mauna Kea, se ha implementado para realizar estudios de microlentes, curvas de luz que se adquirieron y que se han puesto a disposición del público en la página Archivo de exoplanetas de la NASA. (NASA, 2021)

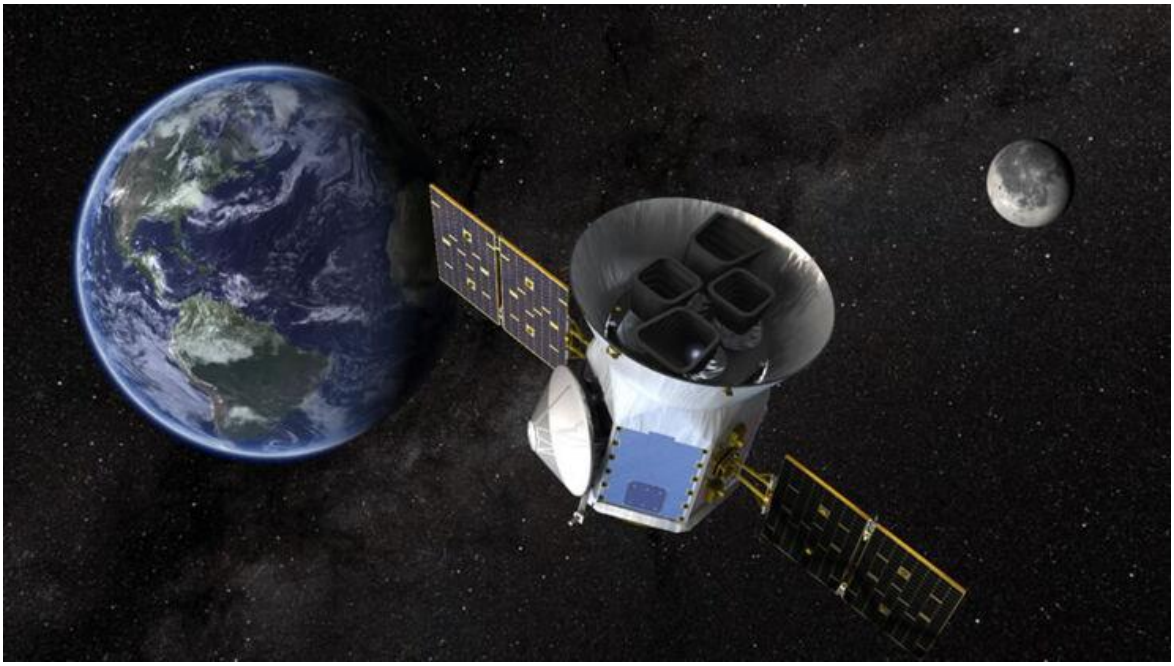
Figura 1-11: Telescopio de exploración en infrarrojo.

Nombre de la fuente: Circulo astronómico. **Recuperado de:** http://www.circuloastronomico.cl/observatorios/observatorios2.html

TESS

El satélite de estudio de exoplanetas en tránsito (TESS) de la NASA de la figura 1-12 inició su misión el 18 de abril del 2018 con el objetivo de descubrir planetas alrededor de estrellas brillantes cercanas. (Exoplanet Exploration, 2021).

Figura 1-12: TESS, el Cazador de Planetas de la NASA.

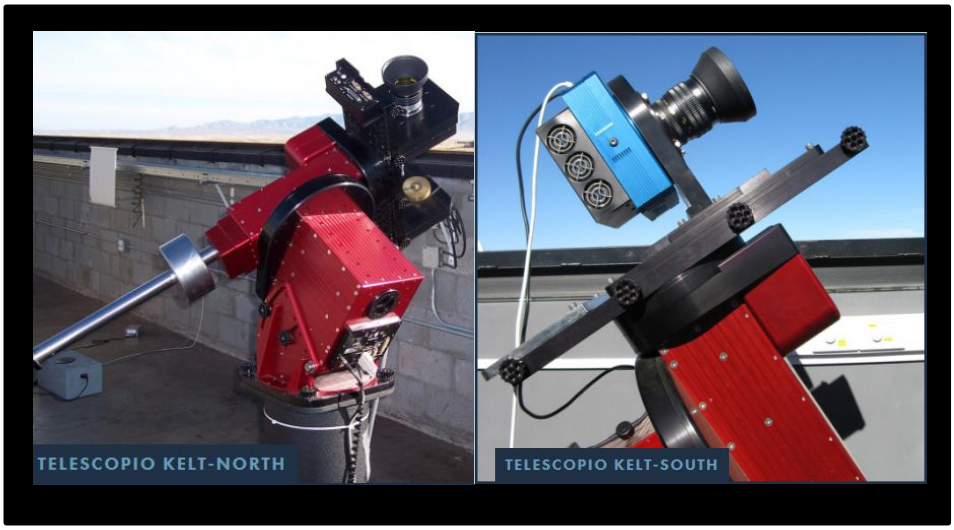


Nombre de la fuente: NASA en español. **Recuperado de:** https://www.lanasa.net/misiones/sondas/tess-el-cazador-de-planetas-de-la-nasa-completa-su-mision-principal

KELT

KELT es un estudio astronómico que descubre planetas fuera del sistema solar empleando dos telescopios robóticos que se observan en la figura 1-13. Cada telescopio escanea el cielo midiendo el brillo de millones de estrellas. Toma imágenes (fotometría) de estrellas cada noche intentando observar los planetas fuera del Sistema solar empleando el método de transito para realizar las mencionadas observaciones. (KELT, 2021).

Figura 1-13: Telescopios robóticos del estudio KELT.



Nombre de la fuente: KELT. **Recuperado de:** https://keltsurvey.org/telescopes.

1.3.2 Métodos de observación

Actualmente existen cinco métodos empleados por los científicos de la NASA para descubrir exoplanetas por medio de las sondas mencionadas, los cuales son: Tránsito y velocidad radial, Imagen directa, microlente gravitacional y astrometría. (Exoplanet Exploration, 2021).

Transito

El método de tránsito de la figura 1-14 se emplea en el momento en que un planeta pasa exactamente entre un observador y la estrella que está orbitando y se genera un bloqueo de la luz estelar desplegada, gracias a esto la luz de la estrella durante un corto periodo de tiempo se vuelve más tenue. Este cambio pequeño que se genera es suficiente para indicar a los astrónomos y observadores sobre la posible presencia de un exoplaneta alrededor de una estrella distante. Empleando este método se han observado 3325 planetas hasta el momento, siendo una de las técnicas principales. (Exoplanet Exploration, 2021).

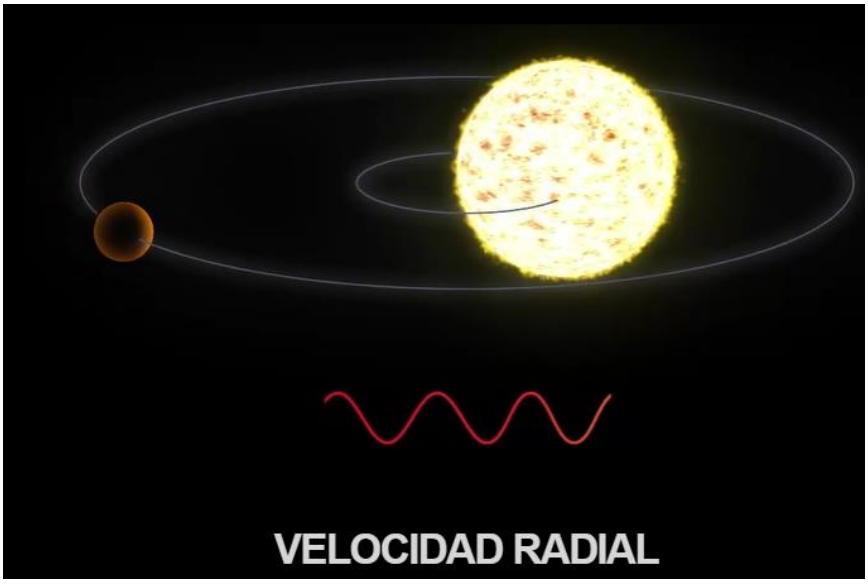
Figura 1-14: Método de tránsito.

Nombre de la fuente: Exoplanet Exploration. **Recuperado de:** https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/2.

Velocidad radial

El método de las velocidades radiales para detectar exoplanetas de la figura 1-15 se emplea al obtener las características y configuración del sistema planetario. Las velocidades radiales de la estrella pueden ser tomadas a través del efecto Doppler en la luz de la estrella y esta velocidad tiene una relación con la velocidad a la que rotan los planetas en torno la estrella.

Figura 1-15: Método de las velocidades radiales.



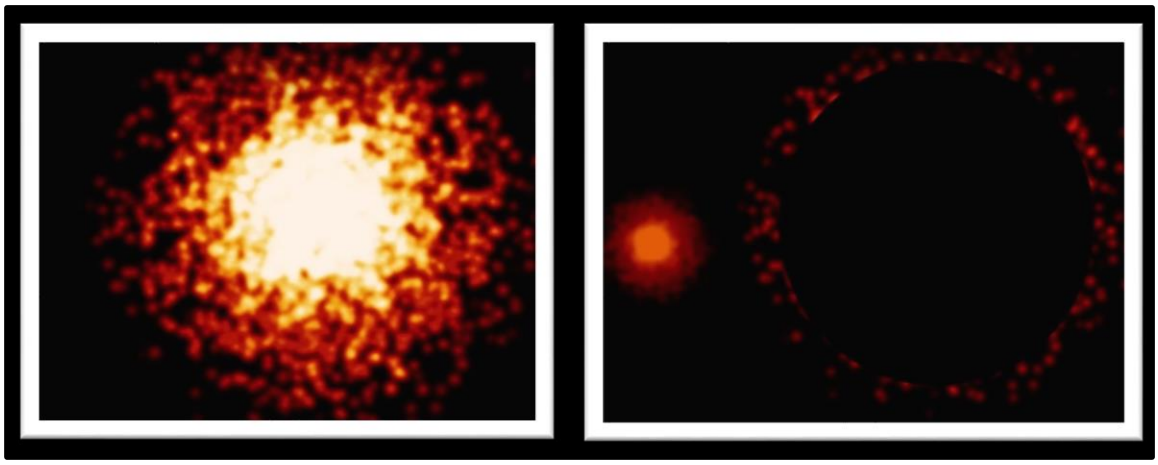
Nombre de la fuente: Exoplanet Exploration. **Recuperado de:** https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/1.

La velocidad radial de un planeta que orbita una estrella que a su vez es orbitada por más planetas es la resultante del efecto conjunto de todos los planetas que se encuentran orbitando dicha estrella. Empleando este método se han observado 837 planetas hasta el momento, siendo una de las técnicas principales junto al método de tránsito. (SINCLAIR, 2016).

Imagen directa

En el método de imagen directa se realiza la toma de fotografías de los exoplanetas como la de la figura 1-16 logradas al reducir el resplandor de las estrellas que orbitan. Una vez que se reduce el resplandor de la estrella, se logra observar mejor los objetos alrededor de la misma y reconocer los posibles exoplanetas que se encuentren a su alrededor. Empleando este método se han observado 51 planetas hasta el momento. (Exoplanet Exploration, 2021).

Figura 1-16: Fotografías empleando el método de imagen directa.

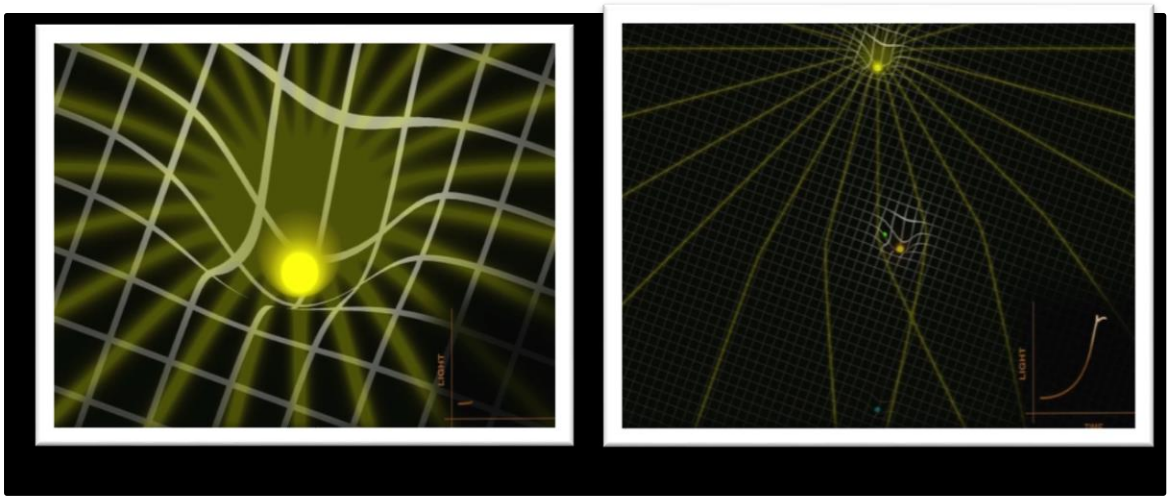


Nombre de la fuente: Exoplanet Exploration. **Recuperado de:** https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/3.

Microlente gravitacional

El efecto de microlente gravitacional de la figura 1-17 se presenta cuando los campos de gravedad del exoplaneta y la estrella que orbita actúan para aumentar o focalizar la luz de una estrella que se encuentra distante, en pocas palabras la gravedad de un gran objeto curva la luz que proviene de objetos distantes y la amplifica, actuando como una lente de aumento. Los tres cuerpos tienen que estar alineados para que el método sea efectivo aunque las detecciones generadas al emplear este método no son repetibles y se requiere estudiar el exoplaneta hallado con algún otro método mencionado. Empleando este método se han observado 108 planetas hasta el momento (Liga Iberoamericana de Astronomía, 2014).

Figura 1-17: Esquema de trayectorias de luz en una lente gravitatoria**.**

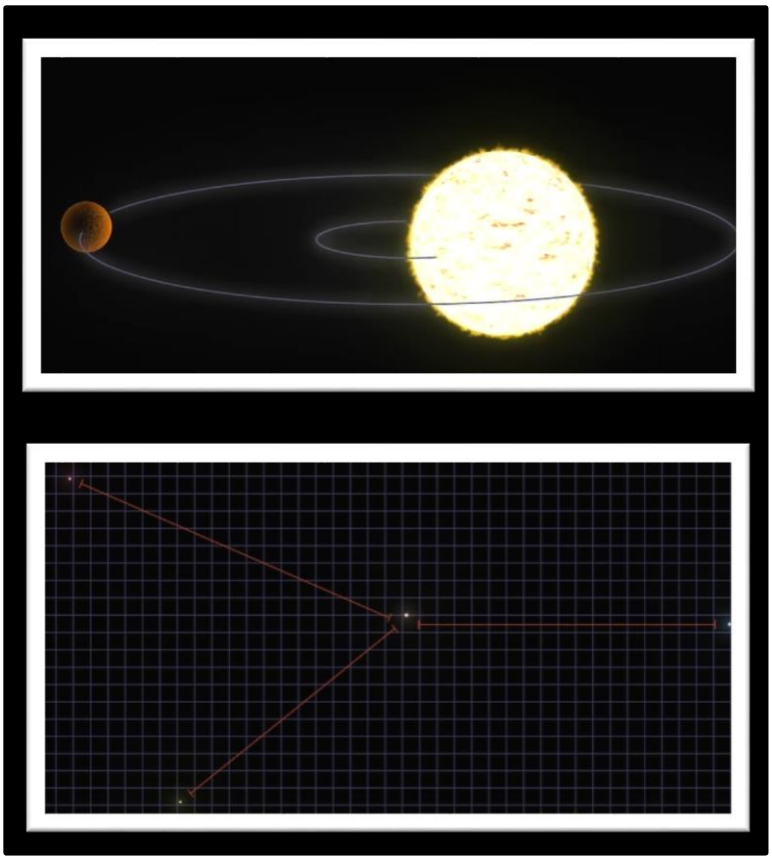


Nombre de la fuente: Exoplanet Exploration. **Recuperado de:** https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/#/4.

Astroemería

El método de la astrometría que se observa en la figura 1-18 es similar al rastreo por velocidad radial mencionado anteriormente. Empleado en la detección de exoplanetas al medir la mínima perturbación regular en la posición de la estrella provocada un cuerpo cercano o posibles exoplanetas que la orbitan. La estrella se mueve en una órbita circular con un radio que depende de la masa del planeta y de su distancia con respecto a la estrella. Empleando este método se ha observado 1 planeta hasta el momento. (Liga Iberoamericana de Astronomía, 2014).

Figura 1-18: Método de la astrometría.



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

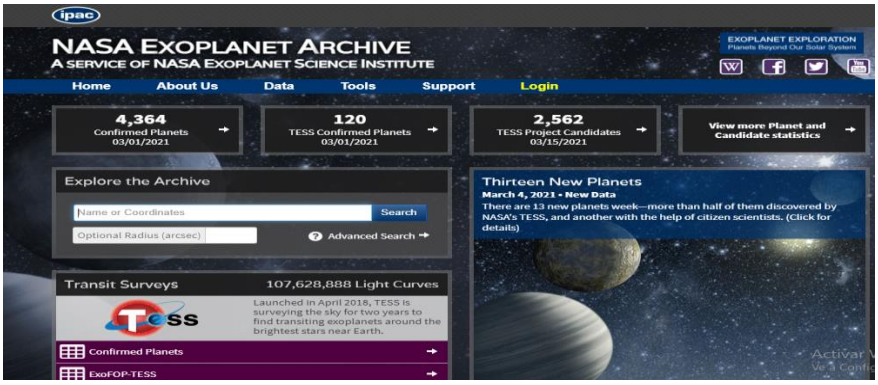
2. Capítulo 2

2.1 Recolección de datos.

El archivo que contiene la información referente a los exoplanetas observados, datos obtenidos a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt, se encuentran publicados en la página de la NASA como un servicio de datos y catálogo astronómico en línea. (NASA, 2021).

El archivo de exoplanetas de la NASA se extrajo de la página que se observa en la figura 2-1, se encuentra contenido en la sección de datos y sistemas planetarios confirmados; contiene información que relaciona directamente la información sobre el exoplaneta y su estrella o estrellas anfitrionas, además de proporcionar gran variedad de herramientas para poder analizar y explorar mejor los datos publicados. La base de datos realiza un reporte de diferentes tipos de datos empleados a lo largo del estudio o búsqueda de los exoplanetas y sus estrellas anfitrionas, datos que permiten caracterizar y clasificar dichos planetas extrasolares en un conjunto de tipos de planetas. Estos datos incluyen parámetros como radios, magnitudes, masas, entre otros datos relevantes que se pueden observar en la tabla 2-1 y tabla 2-2.

Figura 2-1: Página que contiene los archivos de la NASA sobre Exoplanetas.



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

La base de datos contiene 29.318 exoplanetas observados con la información mencionada anteriormente como se observa en la Figura 2-2, de los cuales 4364 son exoplanetas observados y confirmados de la tabla como se ilustra en la Figura 2-3. Cabe aclarar que no todos los exoplanetas tienen todos los datos completos dentro de la base de datos y esto será un factor importante en el procesamiento de dicha información.

Figura 2-2: Base de datos contiene 29.318 datos de observaciones de diferentes observatorios.

```
print(datosExoplanetas1)
```

	pl_name	hostname	pl_letter	...	sy_dist	sy_disterr1	sy_disterr2
0	11 Com b	11 Com	b	...	93.1846	1.92380	-1.92380
1	11 Com b	11 Com	b	...	93.1846	1.92380	-1.92380
2	11 UMi b	11 UMi	b	...	125.3210	1.97650	-1.97650
3	11 UMi b	11 UMi	b	...	125.3210	1.97650	-1.97650
4	11 UMi b	11 UMi	b	...	125.3210	1.97650	-1.97650
...	...	...	...	...	...	...	...
29313	ups And d	ups And	d	...	13.4054	0.06350	-0.06290
29314	ups And d	ups And	d	...	13.4054	0.06350	-0.06290
29315	ups And d	ups And	d	...	13.4054	0.06350	-0.06290
29316	xi Aql b	xi Aql	b	...	56.1858	0.55975	-0.55975
29317	xi Aql b	xi Aql	b	...	56.1858	0.55975	-0.55975

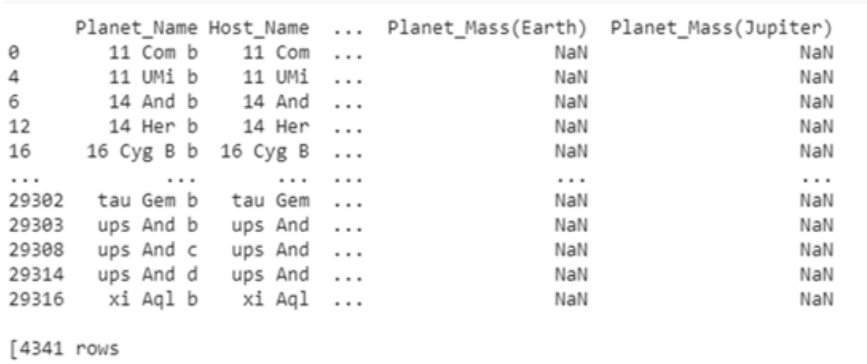
[29318 rows x 42 columns]

Nombre de la fuente: The NASA Exoplanet Archive. Recuperado de:

https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

Figura 2-3: Base de datos 4341 datos de observaciones de diferentes observatorios.

.



	Planet_Name	Host_Name	...	Planet_Mass(Earth)	Planet_Mass(Jupiter)
0	11 Com b	11 Com	...	NaN	NaN
4	11 UMi b	11 UMi	...	NaN	NaN
6	14 And b	14 And	...	NaN	NaN
12	14 Her b	14 Her	...	NaN	NaN
16	16 Cyg B b	16 Cyg B	...	NaN	NaN
...	...	...	...	...	...
29302	tau Gem b	tau Gem	...	NaN	NaN
29303	ups And b	ups And	...	NaN	NaN
29308	ups And c	ups And	...	NaN	NaN
29314	ups And d	ups And	...	NaN	NaN
29316	xi Aql b	xi Aql	...	NaN	NaN

[4341 rows

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

Con respecto a la clasificación de los planetas extrasolares, los científicos hasta el momento han categorizado los exoplanetas de la siguiente manera: gigante gaseoso, neptuniano, supertierra y terrestre.

Cada tipo de planeta que se mencionó, presenta variaciones en apariencia interior y exterior dependiendo de su composición según el tipo en el que se clasifique.

El tamaño y la masa del planeta son datos indispensables para poder realizar una correcta clasificación de dichos planetas. El “valle del radio” o más conocida como la brecha de Fulton, en honor a su principal autor Benjamin Fulton nos habla de planetas con tamaños entre 1,5 y 2 veces el tamaño (diámetro) de la tierra, planetas que son muy raros y que se clasificarían como súper-tierras. (Exoplanet Exploration, 2021).

Tabla 2-1: Documentación de datos astronómicos sobre exoplanetas obtenidos a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt.

.

Nombre de la columna de la base de datos	Etiqueta de mesa	Descripción	Columna de incertidumbres positivo(+) negativo(-)	Columna de limite
pl_name	Nombre del planeta	Nombre del planeta más comúnmente empleado en la literatura.		
hostname	Nombre de host	Nombre estelar más utilizado en la literatura.		
pl_letter	Letra planeta	Letra asignada al componente planetario de un sistema planetario.		
default_flag	Conjunto de parámetros predeterminados	Bandera booleana que indica si se ha seleccionado un conjunto dado de parámetros planetarios por defecto(1=si, 0=no)		
sy_snum	Número de estrellas	Número de 4 estrellas en el sistema planetario.		
discoverymethod	Método de descubrimiento	Método por el cual se identificó el planeta por primera vez.		
pl_orbper	Periodo orbital (días)	Tiempo que tarda el planeta en hacer una órbita completa alrededor de la estrella o sistema anfitrión.	(+) pl_orbperr1 (-) pl_orbperr2	pl_orbperlim
pl_rade	Radio del planeta (radio de la Tierra)	Longitud de un segmento de línea desde el centro del planeta hasta su superficie, medida en unidades de radio de la tierra.	(+) pl_radeerr1 (-) pl_radeerr2	pl_radelim

pl_radj	Radio del planeta (radio de Júpiter)	Longitud de un segmento de línea desde el centro del planeta hasta su superficie, medida en unidades de radio de Júpiter.	(+) pl_radjerr1 (-) pl_radjerr2	pl_radjlim
pl_masse	Masa del planeta (masa de la Tierra)	Cantidad de materia contenida en el planeta, medida en unidades de masa de la tierra.	(+) pl_masseerr1 (-) pl_masseerr2	pl_masse lim

Tabla 2-2: Documentación de datos astronómicos sobre exoplanetas obtenidos a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt.

.

Nombre de la columna de la base de datos	Etiqueta de mesa	Descripción	Columna de incertidumbres positivo (+) Negativo (-)	Columna de limite
pl_massj	Masa del planeta (masa de Júpiter)	Cantidad de materia contenida en el planeta medida en unidades de masa de Júpiter.	(+) pl_massjerr1 (-) pl_massjerr2	pl_massjlim
pl_dens	Densidad planetaria en(g/cm^3)	Cantidad de masa por unidad de volumen del planeta.	(+) pl_denserr1 (-) pl_denserr2	pl_denslim
st_rad	Radio estelar (radio solar)	Longitud de un segmento de línea desde el centro de la estrella hasta su superficie, medida en unidades de radio del sol.	(+) st_raderr1 (-) st_raderr2	st_radlim
st_mass	Masa estelar (masa solar)	Cantidad de materia contenida en la estrella, medida en unidades de masa del sol.	(+) st_masserr1 (-) st_masserr2	st_masslim
sy_dist	Distancia (pc)	Distancia al sistema planetario en unidades de parsecs.	(+) sy_disterr1 (-) sy_disterr2	

La segunda base de datos se creó extrayendo la información contenida de una enciclopedia exoplanetaria que mantiene en continua actualización y que posee datos detallados sobre todos los exoplanetas conocidos hasta el momento como se observa en la Figura 2-4 (Exoplanet Exploration, 2021).En esta base de datos se encuentran solo los planetas confirmados a diferencia de la primera y solo posee información del nombre del exoplaneta, años luz de la tierra, masa del planeta, magnitud estelar, la fecha de descubrimiento y el tipo de planeta, ver Tabla 2-3.

Figura 2-4: Base de datos con los exoplanetas confirmados y su clasificación por tipo de planeta.

NAME ↑	LIGHT-YEARS FROM EARTH	PLANET MASS	STELLAR MAGNITUDE	DISCOVERY DATE
11 Comae Berenices b	304	19.4 Jupiters	4.72307	2007
11 Ursae Minoris b	409	14.74 Jupiters	5.013	2009
14 Andromedae b	246	4.8 Jupiters	5.23133	2008
14 Herculis b	58	4.66 Jupiters	6.61935	2002
16 Cygni B b	69	1.78 Jupiters	6.215	1996
18 Delphini b	249	10.3 Jupiters	5.51048	2008
1RXS J160929.1-210524 b	454	8 Jupiters	12.618	2008
24 Bootis b	313	0.91 Jupiters	5.59	2018
24 Sextantis b	235	1.99 Jupiters	6.4535	2010
24 Sextantis c	235	0.86 Jupiters	6.4535	2010

1 of 438 Back to top

Nombre de la fuente: Exoplanet Exploration. **Recuperado de:** https://exoplanets.nasa.gov/discovery/exoplanet-catalog/

Estas dos bases de datos unidas presentan toda la información recolectada hasta el momento para el estudio de los planetas Extrasolares y será la información inicial con la que se partirá.

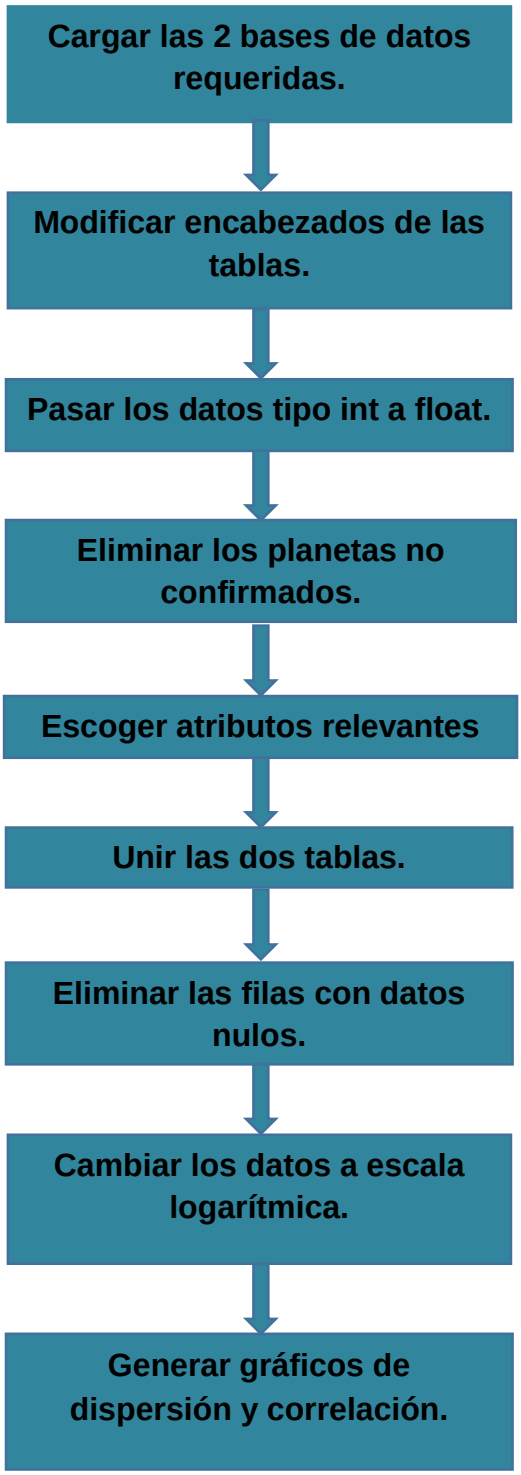
Tabla 2-3: Información contenida de la enciclopedia Exoplanetaria obtenida a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt.

Nombre de la columna de la base de datos	Descripción
Años luz de la tierra	Distancia con respecto a la tierra que la luz recorre en un año terrestre.
Masa del planeta	Cantidad de materia contenida en el planeta, medida en unidades de masa de la tierra.
Magnitud estelar	Medida utilizada para medir el brillo aparente de las estrellas.
Fecha de descubrimiento	Año en el que se descubrió o se confirmó el ExoPlaneta.
Tipo del planeta	Clasificación según sus características.

2.2 Limpieza, filtrado y preprocesamiento de datos.

Es de vital importancia realizar una limpieza, filtrado y un preprocesamiento de la información trabajada buscando sacar o eliminar todos los datos atípicos encontrados sin afectar toda la información, rectificar si la información está normalizada y rectificar la escala en la que se encuentre la información para de esta manera crear un modelado confiable del problema. A continuación en la figura 2-5 se observa el diagrama de bloques con los pasos involucrados en este proceso.

Figura 2-5: Diagrama de bloques de limpieza, filtrado y preprocesamiento de la información**.**



Partiendo de las bases de datos en la figura 2-2, figura 2-3 y figura 2-4 extraídas de páginas oficiales de la NASA, donde cada tabla presenta diferentes atributos relevantes de estudio que permiten realizar la clasificación de los exoplanetas en los distintos tipos de planetas según las características que poseen cada uno. Se inicia con la exploración de la información contenida en las tablas para una posterior reorganización y renombramiento de las columnas que permita reconocer de una forma inmediata el tipo de dato contenido en cada columna, siendo importante trabajar con archivos con una extensión .csv para ser leído sin ningún problema por el software de Python como un data frame como se observa en la figura 2-6.

Figura 2-6: Base de datos con nombres modificados y detallados de las columnas.

	Planet_Name	Host_Name	...	Distance_Upper_Unc(pc)	Distance_Lower_Unc(pc)
0	11 Com b	11 Com	...	1.92380	-1.92380
1	11 Com b	11 Com	...	1.92380	-1.92380
2	11 UMi b	11 UMi	...	1.97650	-1.97650
3	11 UMi b	11 UMi	...	1.97650	-1.97650
4	11 UMi b	11 UMi	...	1.97650	-1.97650
...	...	...	...	...	...
29313	ups And d	ups And	...	0.06350	-0.06290
29314	ups And d	ups And	...	0.06350	-0.06290
29315	ups And d	ups And	...	0.06350	-0.06290
29316	xi Aql b	xi Aql	...	0.55975	-0.55975
29317	xi Aql b	xi Aql	...	0.55975	-0.55975

[29318 rows x 42 columns]

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

Se busca conocer el tipo de dato almacenado en cada columna y estandarizar la información para que a futuro cuando se requiera realizar operaciones entre columnas o relacionar la información de alguna forma para generar gráficas que permitan extraer información importante de la base de datos, no se generen conflictos. Además, se requiere unificar las dos tablas extraídas. La primera tabla contiene 29367 planetas confirmados y sin confirmar. Un exoplaneta sin confirmar o candidato es un planeta probable descubierto por un telescopio, pero aún no se ha demostrado que exista realmente (Exoplanet Exploration, 2021). Este parámetro de confirmación se define por la columna

“Default_Parameter_Set”. Cuando el parámetro “Default_Parameter_Set” sea igual a 1, el Exoplaneta observado es un planeta Extrasolar confirmado, como se observa en la figura 2-7.

Figura 2-7: Base de datos extraída de los archivos de la nasa con el parámetro Default de confirmación del Exoplaneta observado.

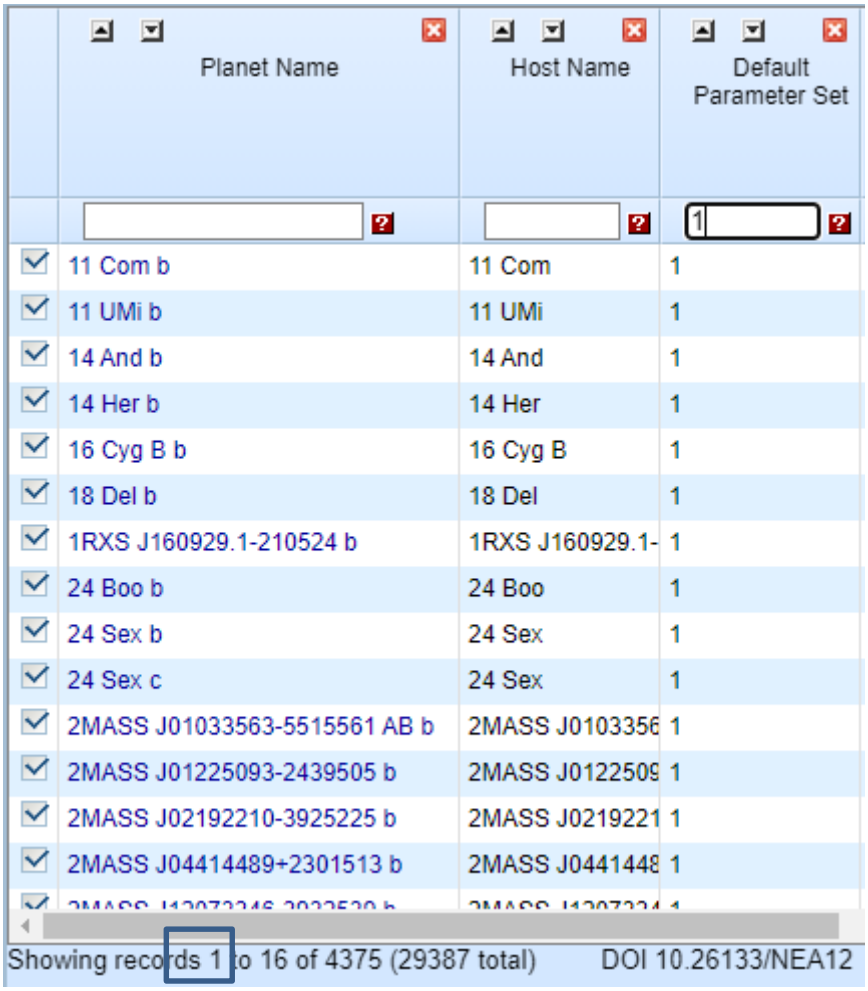
Planet Name	Host Name	Default Parameter Set
11 Com b	11 Com	0
11 Com b	11 Com	1
11 UMi b	11 UMi	0
11 UMi b	11 UMi	1
11 UMi b	11 UMi	0
14 And b	14 And	1
14 And b	14 And	0

Nombre de la fuente: The NASA Exoplanet Archive. Recuperado de:

https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS

Ya que se requiere trabajar solo con los exoplanetas confirmados y clasificados, se eliminan de la primera tabla los planetas sin confirmación. Al eliminar en la primera base de datos todos los planetas que no han sido confirmados (Default_Parameter_Set = 0), la base se reduce a 4375 planetas confirmados, como se observa en la figura 2-8.

Figura 2.8: Base de datos extraída de los archivos de la NASA con el parámetro Default de confirmación igual a 1 (solo exoplanetas confirmados).



Planet Name	Host Name	Default Parameter Set
11 Com b	11 Com	1
11 UMi b	11 UMi	1
14 And b	14 And	1
14 Her b	14 Her	1
16 Cyg B b	16 Cyg B	1
18 Del b	18 Del	1
1RXS J160929.1-210524 b	1RXS J160929.1-	1
24 Boo b	24 Boo	1
24 Sex b	24 Sex	1
24 Sex c	24 Sex	1
2MASS J01033563-5515561 AB b	2MASS J0103356	1
2MASS J01225093-2439505 b	2MASS J0122509	1
2MASS J02192210-3925225 b	2MASS J0219221	1
2MASS J04414489+2301513 b	2MASS J0441448	1

Showing records 1 to 16 of 4375 (29387 total) DOI 10.26133/NEA12

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

A continuación se busca reducir aún más la cantidad de datos existentes dentro de las tablas, para esto se requiere definir cuáles serán los parámetros o atributos determinantes o indispensables a la hora de realizar la clasificación de cada exoplaneta.

El tamaño (Radio del planeta) y la masa del planeta son datos indispensables para poder realizar una correcta clasificación de dichos planetas como se mencionó anteriormente, por ese motivo se conservan las mencionadas columnas con los datos existentes con respecto a la Tierra de referencia para estandarizar la información, además de conservar

el tipo de planeta, que se empleara como etiqueta para los algoritmos de aprendizaje supervisado que se implementen.

El periodo orbital es una columna que se debe incluir para obtener un resultado más natural con menos datos atípicos. Entre más características relevantes se incluyen se obtendrá un mejor gráfico con una segmentación de la información más definida con respecto a la clasificación de los planetas que se conocen.

La figura 2-9 ilustra una parte de la tabla con los atributos relevantes que se emplearan para la clasificación de los exoplanetas observados.

Figura 2-9: Atributos relevantes seleccionados para la clasificación de Exoplanetas**.**

Index	Planet_Name	Planet_Radius(Earth)	Orbital_Period(Days)	Planet_Mass_Earth
16	2MASS J21402931+1625183 A b	10.31	7336.5	6662.1
51	BD+20 594 b	2.578	41.6855	22.2481
68	CoRoT-1 b	16.7	1.50896	327.54
69	CoRoT-10 b	10.87	13.2406	874.5

Nombre de la fuente: Spyder(Python 3.8).

En este punto las dos bases de datos ya contienen el mismo número de filas o exoplanetas y se pueden unificar para de esta forma conseguir una sola tabla con toda la información indispensable brindada por la NASA para la clasificación de los mencionados planetas. La tabla resultante de la extracción de los mencionados atributos, la unificación de las tablas y la eliminación de las filas con datos desconocidos o nulos se observa en la figura 2-10, donde se tiene una matriz de 5 columnas o atributos que son: El nombre del planeta, la masa del planeta con respecto a la tierra, el radio del planeta con respecto a la tierra, el periodo orbital en días y el tipo de planeta.

Figura 2-10: Base de datos resultante con los atributos finales seleccionados**.**

	Planet_Name	...	Planet_Type
16	2MASS J21402931+1625183 A b	...	Gas Giant
51	BD+20 594 b	...	Neptune-like
68	CoRoT-1 b	...	Gas Giant
69	CoRoT-10 b	...	Gas Giant
70	CoRoT-11 b	...	Gas Giant
...	...	...	...
4296	XO-3 b	...	Gas Giant
4297	XO-4 b	...	Gas Giant
4298	XO-5 b	...	Gas Giant
4299	XO-6 b	...	Gas Giant
4300	XO-7 b	...	Gas Giant

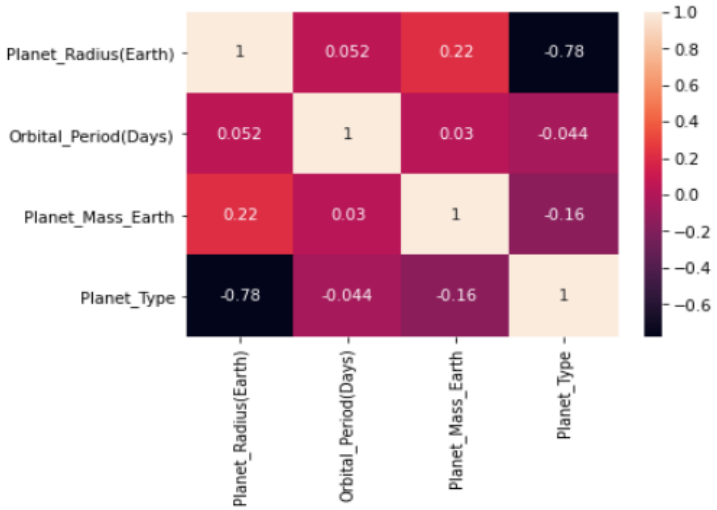
[3286 rows x 5 columns]

Nombre de la fuente: Spyder(Python 3.8).

La correlación de los datos es fundamental a la hora de poder definir que parámetros escoger, buscando emplear los datos y la escala de los mismos que genere una mayor correlación. En la figura 2-11 se observa la matriz de correlación de la tabla resultante con los datos finales a emplear en la aplicaciones de los algoritmos de Machine learning, donde 1.0 indica la mayor correlación existente.

Inicialmente se realiza la matriz de correlaciones con los datos iniciales figura 2-11 y posteriormente se realiza la misma matriz de correlación pero empleando los datos en escala logarítmica figura 2-12.

Figura 2-11: Matriz de correlaciones de la tabla final sin modificar los datos.

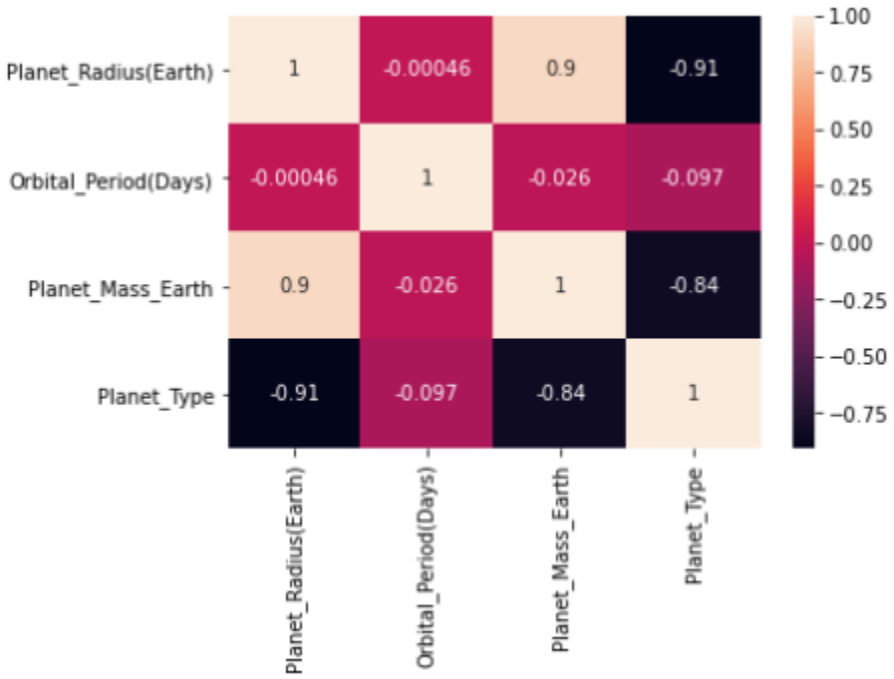


Nombre de la fuente: Spyder(Python 3.8).

La correlación es mayor al emplear los datos en una escala logarítmica como se puede observar en la figura 2-12 donde se grafica la matriz de correlaciones con los datos contenidos en la tabla en escala logarítmica, por este motivo las gráficas de dispersión de la figura 2-13, figura 2-14 y la figura 2-15 se generaron en dicha escala y se aplicarán de esta forma en los algoritmos de aprendizaje supervisado y no supervisado buscando mejorar la precisión y exactitud del modelo predictivo generado con estas técnicas de aprendizaje automático como se presenta en el capítulo siguiente.

En la figura 2-12 se observa que los datos del radio de planeta en función del radio de la tierra y la masa del planeta en función a la masa de la tierra tienen una correlación de 0.9, siendo la correlación más alta obtenida y el motivo principal de seleccionar los mencionados atributos como atributos relevantes.

Figura 2-12: Matriz de correlaciones de la tabla final con los datos en escala logarítmica.

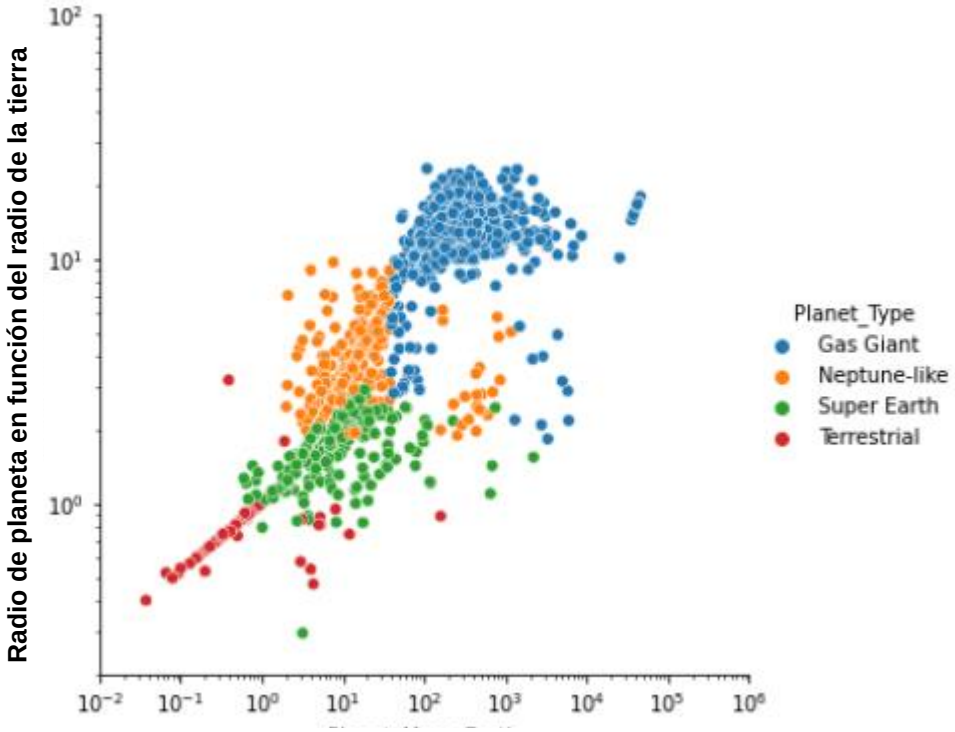


Nombre de la fuente: Spyder(Python 3.8).

La cuestión estadística con respecto a la graficaciòn y relación de las columnas escogidas es fundamental a la hora de observar el comportamiento de los datos y conseguir moldearlos de una forma confiable, buscando resultados más naturales y linealmente

separables con respecto a su distribución en el plano. En este punto al generar los gráficos de dispersión que se observan en la figura 2-13, figura 2-14 y la figura 2-15 que relacionan los atributos seleccionados, se observa datos sobrelapados, por este motivo se requiere eliminar estos datos atípicos sin afectar toda la información o definir si estos datos no afectarán en gran medida la exactitud del modelo generado al aplicar los algoritmos.

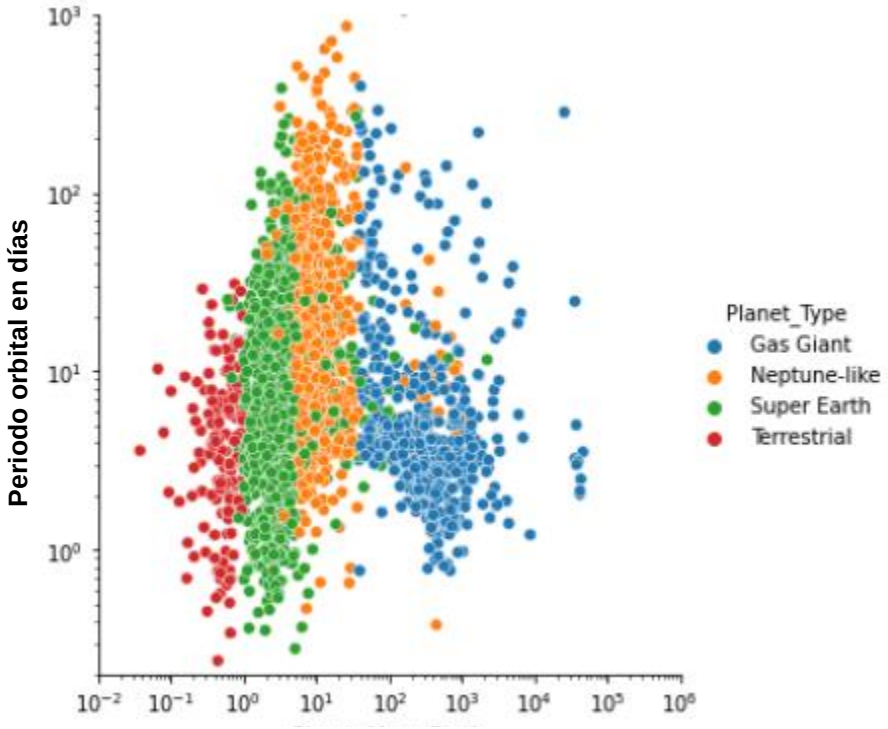
Figura 2-13: Gráfica de dispersión del radio del planeta con respecto a la masa del planeta en función de la tierra.



Nombre de la fuente: Spyder(Python 3.8).

En este punto se observan menos datos atípicos en la figura 2-13, figura 2-14 y la figura 2-15 y se puede observar una buena separación entre los grupos de datos o tipos de planetas.

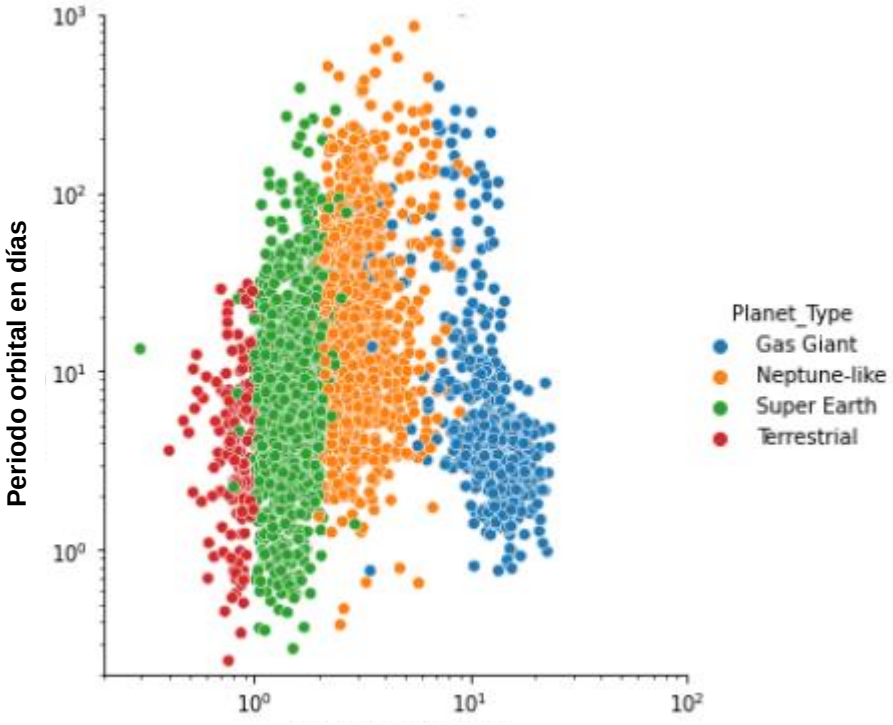
Figura 2-14: Gráfica de dispersión del periodo orbital en días con respecto a la masa del planeta en función de la tierra.



Nombre de la fuente: Spyder(Python 3.8).

El solapamiento que se observa en la figura 2-13, figura 2-14 y la figura 2-15 de algunos datos o datos atípicos que se observan en mayor instancia son los planetas tipo Neptuno y los super tierra, debido a que existen planetas clasificados como tipo Neptuno que se encuentran en el rango de las super tierra ya que comparten muchas características entre sí.

Figura 2-15: Gráfica de dispersión del periodo orbital en días con respecto al radio del planeta en función de la tierra.



Nombre de la fuente: Spyder(Python 3.8).

Lo importante a resaltar en los gráficos de la figura 2-13, figura 2-14 y la figura 2-15, es que a pesar del pequeño grupo de datos atípicos que se observan, existen un agrupamiento lo suficientemente remarcado como para poder diferenciar los 4 tipos de planetas que se requieren clasificar empleando las técnicas y algoritmos de machine Learning.

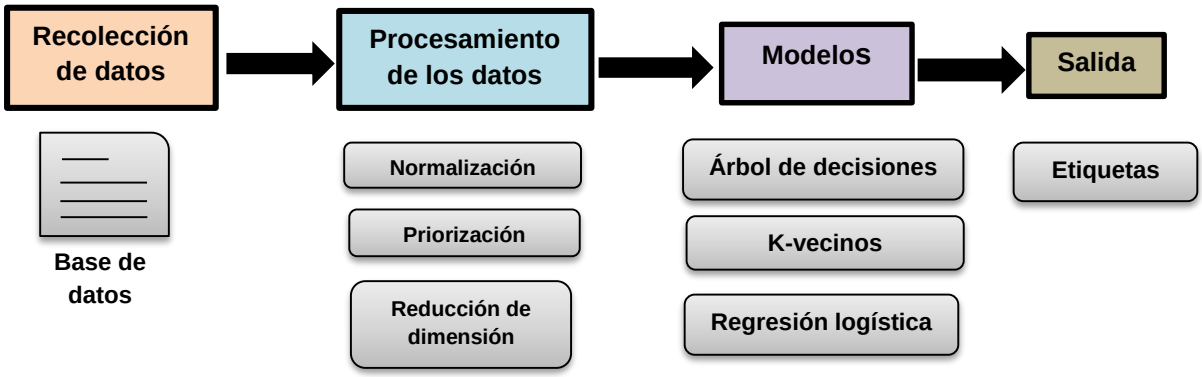
3. Capítulo 3

3.1 Metodología.

3.1.1 Técnicas de Aprendizaje Supervisado.

Para realizar el presente trabajo que tiene como principal objetivo implementar técnicas y algoritmos de aprendizaje supervisado y no supervisado de ML en el análisis de datos astronómicos sobre exoplanetas recolectados a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt , se ha implementado la siguiente metodología ilustrada en la figura 3-1 para los algoritmos de aprendizaje supervisado.

Figura 3-1: Esquema de la metodología implementada para los algoritmos de MLsupervisados.

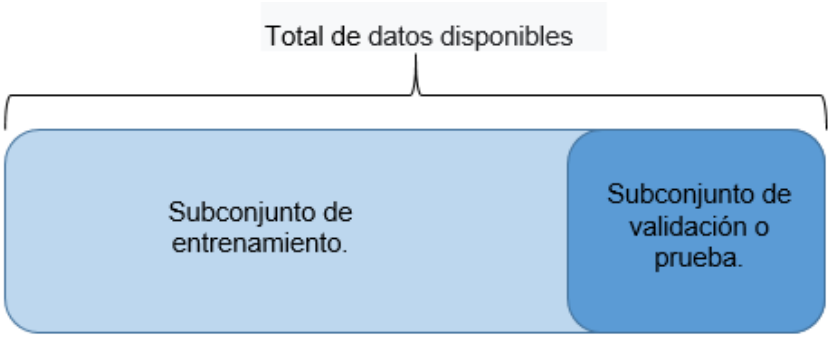


Inicialmente se requiere codificar la información contenida en la columna correspondiente a tipo de planeta que se definió inicialmente como la etiqueta que se entregara a los algoritmos de aprendizaje supervisado, debido a que se requiere tener las variables

categóricas como variables numéricas y evitar conflictos a la hora de realizar relaciones entre columnas en el modelado predictivo.

A continuación se realiza la división de la base de datos en dos conjuntos como se observa en la figura 3-2. El primer subconjunto de datos es el de entrenamiento, empleado como su nombre lo dice para entrenar el algoritmo. El segundo conjunto es el de prueba o validación, que representa el subconjunto de datos empleados para probar el algoritmo después de ser entrenado. El subconjunto de validación debe ser lo suficientemente grande para generar buenos resultados estadísticos y representativos, esto quiere decir que debe compartir características relevantes con el subconjunto de datos de entrenamiento definido, todo esto con el fin de generar un modelo predictivo que generalice correctamente los datos nuevos ingresados al algoritmo. (Fuentes, 2019).

Figura 3-2: Separación de datos en subconjuntos de entrenamiento y validación.



Para continuar, como no se tiene la misma escala en toda la base de datos, se estandarizan o se ajustan las escalas para obtener mejores resultados.

En este punto se requiere implementar el algoritmo de aprendizaje supervisado. Para el algoritmo de árbol de decisión, se toma en consideración todas las observaciones realizadas sobre los exoplanetas para predecir su tipo de planeta y clasificarlos correctamente.

En los árboles de decisión o de clasificación, los nodos representan cada dato que se posee. Cada ramificación contiene los atributos o reglas de clasificación asociadas a una etiqueta o tipo de planeta. Estas reglas de decisión, se pueden expresar de la siguiente manera: “Si... entonces...". Cada atributo representativo implementado en el modelo,

ayuda a que el modelo realice generalizaciones o predicciones más precisas. (LucidChart, 2021)

Para el algoritmo de k vecinos más cercanos simplemente se busca en las observaciones que se tengan más cercanas a la que se está intentando etiquetar o clasificar y realiza la clasificación del punto de interés apoyado en la mayoría de datos que le rodean. Primero, se Calcula la distancia entre el dato que se desea clasificar y el resto de datos dentro del subgrupo de entrenamiento definido. Luego, selecciona los k elementos más cercanos. Por último, se define la etiqueta que se encuentre con más recurrencia o la predominante y la emplea para la clasificación final del punto de interés. (Bagnato, 2018).

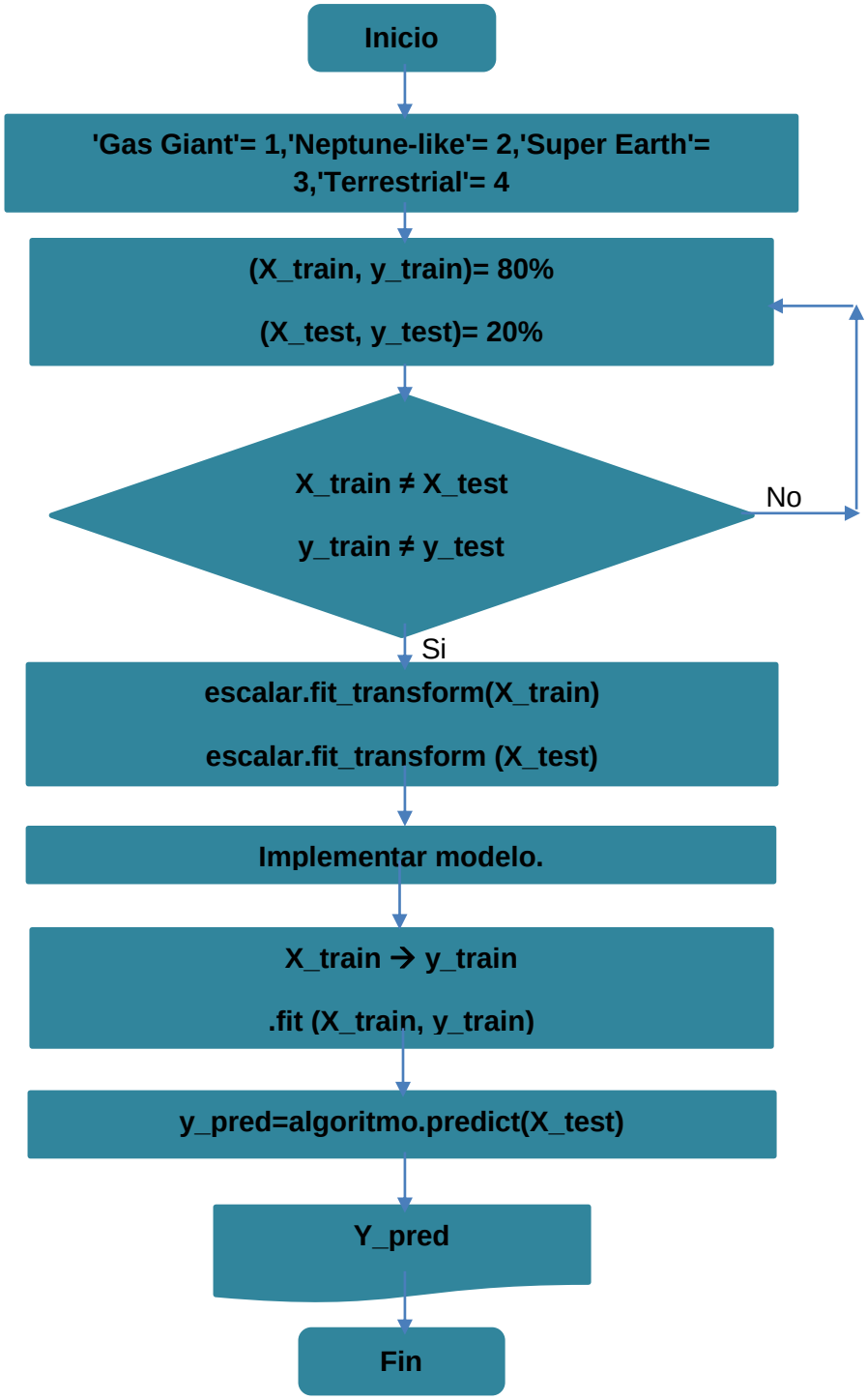
Para el algoritmo de regresión logística se comienza estudiando la probabilidad de que la variable dependiente(x) tenga el valor de 1,2,3 o 4 en función de un conjunto de variables independientes o predictores, que sería la variable etiqueta que define el tipo de planeta al que pertenece el nuevo dato ingresado al algoritmo. Las probabilidades que describen el posible resultado predictivo o tipo de planeta al que pertenece el dato de interés que se obtenga de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística. (Mercado, 2017).

Después de estandarizar la información e implementar el algoritmo a implementar, se realiza el entrenamiento del modelado a través del método Fit que recibe los valores de entrenamiento de X (variable dependiente) y Y (variable independiente).En el proceso de entrenamiento se le entrega al algoritmo los ejemplos o respuestas que le permitirán realizar las generalizaciones necesarias con los datos nuevos ingresados al modelo predictivo y conseguir clasificar la información correctamente. Posteriormente, se realiza las predicciones del subconjunto de prueba definido con anterioridad.

En la figura 3-3 se encuentra el diagrama de flujo de la técnica de aprendizaje supervisado. En este diagrama de realiza un remplazo de los valores tipo objeto por valores numéricos para emplear como etiquetas. Posteriormente, se divide la base de datos en datos de entrenamiento y datos de prueba, donde los datos de entrenamiento corresponden al 80% de la base de datos y los de prueba al 20% restante, recordando que este grupo de datos deben ser distintos. A continuación se escalan los datos y se implementa el algoritmo supervisado. Por último, se realiza las predicciones con los datos de prueba y se imprimen

los valores obtenidos para corroborar si la generalización o predicción del modelo fue acertada.

Figura 3-3: Diagrama de flujo aprendizaje supervisado.



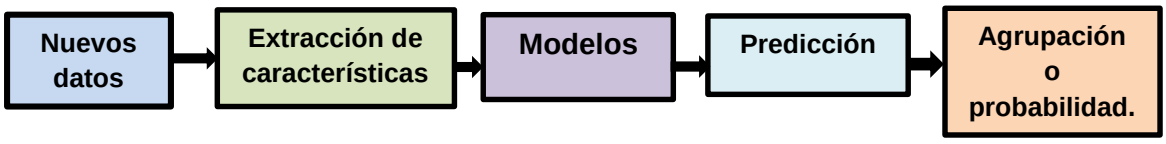
Por último, se requiere evaluar el rendimiento del modelo predictivo generado y para esto existen diversas métricas mencionadas anteriormente. Se debe revisar si los datos que se tienen están balanceados o no, ya que si se encuentran balanceados solo con la métrica de precisión se puede evaluar de manera correcta el rendimiento del mismo de lo contrario, se debería implementar diversas métricas para corroborar su rendimiento.

En términos de clasificación, un conjunto de datos o información que presente una distribución desigual entre sus clases o etiquetas, se puede considerar una base de datos desbalanceada. (Osorio, 2019).

3.1.2 Técnicas de Aprendizaje No Supervisado.

Para el caso de los algoritmos de aprendizaje no supervisado, no se posee una etiqueta o respuesta que le permita aprender al modelo que resultado se desea obtener. En estos algoritmos se elimina la necesidad de entrenamiento previo y la necesidad de clasificación o etiquetamiento manual de casos de entrenamiento como se observa en la figura 3-4.

Figura 3-4: Diagrama de aprendizaje no supervisado.



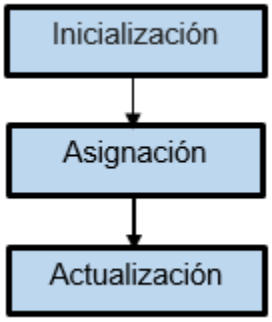
Para los algoritmos de aprendizaje no supervisado no existe un esquema general bien definido para guiar su implementación .Este algoritmo solo obtiene información de la base de datos entregada inicialmente y busca las semejanzas en los mismos para extraer patrones que le permita realizar una correcta generalización. Algunas técnicas que emplean estos algoritmos son: Clusterización (agrupación) y jerarquía.

La agrupación es la técnica que agrupa automática los datos ingresados. Como no se posee una respuesta o etiqueta, la evaluación o clasificación de los datos identificados, es un poco subjetiva a diferencia de las técnicas de aprendizaje supervisado. Estas técnicas intentan descubrir cuál es la mejor forma de agrupar la información ingresada y algunas requieren que se especifique el número de clases o grupos que se requiere generar. (Heras, 2020)

Un algoritmo de agrupación a implementar es k-means. Para el algoritmo k-means se requiere especificar el número de grupos que desean encontrar. A este número de grupos se le denomina k. (Heras, 2020)

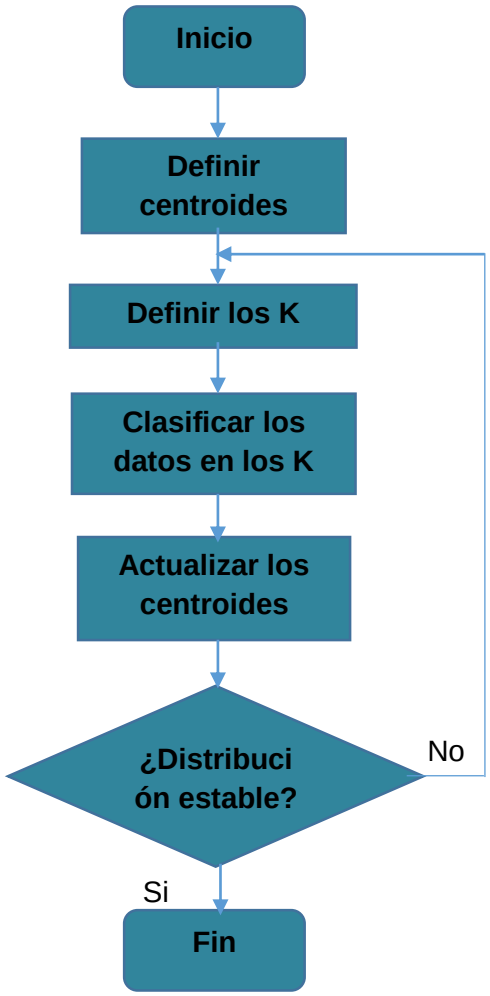
Se ha implementado la siguiente metodología ilustrada en la figura 3-5 para este algoritmo de aprendizaje no supervisado.

Figura 3-5: Diagrama de bloques de la Metodología implementada para el algoritmo k-media o k-means.



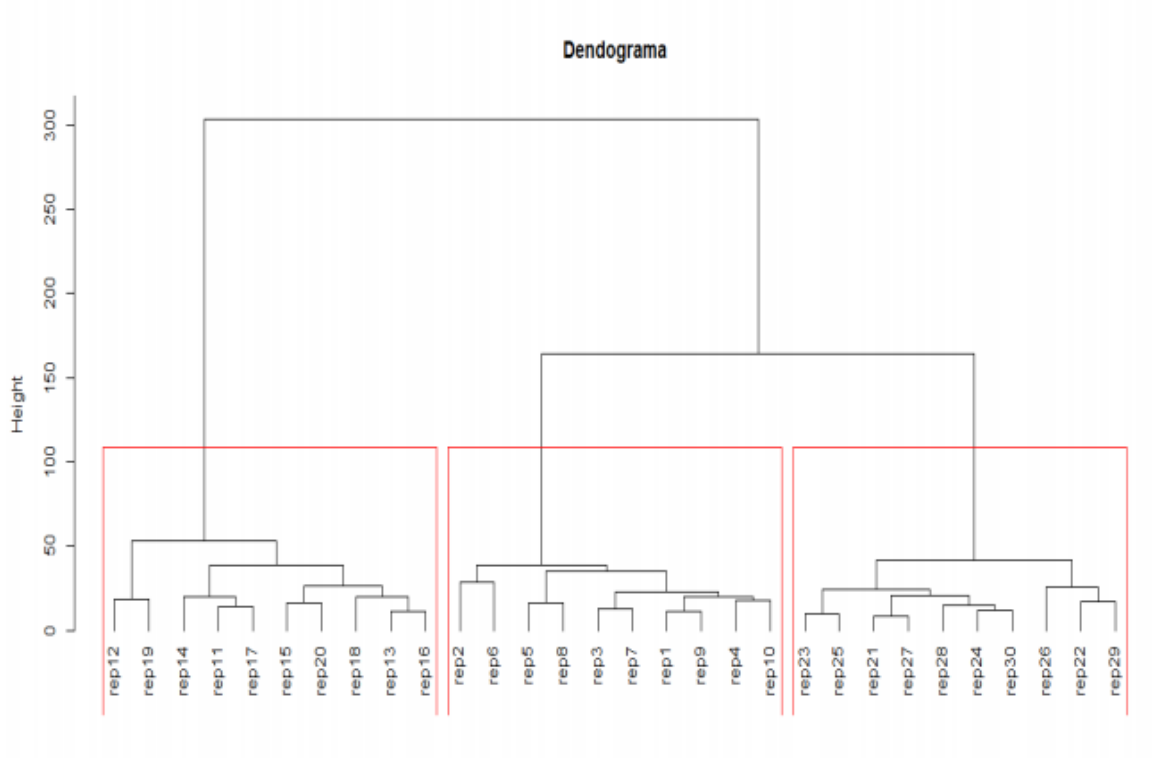
En la figura 3-6 se observa el diagrama de flujo del algoritmo k-means. Primero, se inicializa el algoritmo definiendo la localización de los centroides de los k grupos requeridos aleatoriamente. A continuación se asigna cada dato ingresado al centroide más cercano que encuentre de tal manera que un dato solo puede pertenecer a un clúster o grupo y por último se reacomoda el centroide empleando la media aritmética de las posiciones de todos los datos asignados al grupo que pertenece dicho centriode. (Heras, 2020).

Figura 3-6: Diagrama de flujo del algoritmo k-vecinos más cercanos.



Para el algoritmo de clustering jerárquico se inicia con tantas clases como elementos o planetas se tengan y la distancia establecida entre las clases o grupos es la distancia entre los datos originales. Segundo, se selecciona los elementos más cercanos en la matriz de distancia y se forma una nueva clase con ellos. Tercero, se sustituyen los elementos empleados, modificando la matriz distancia y por último se realiza una iteración constante entre los pasos anteriores, sin incluir el primero para obtener al final todos los planetas agrupados o etiquetados en un único grupo. (Venegas & Pineda Rios, 2017).

Figura 3-7: Grafico del dendograma.

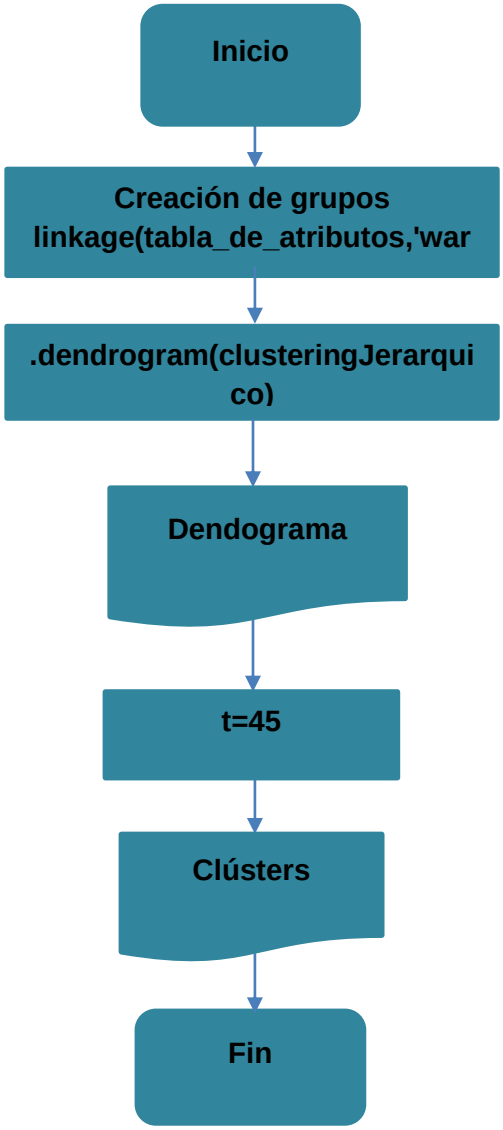


Nombre de la fuente: Método de Cluster Jerárquicos para Datos Funcionales Multivariados: Una Aplicación en Mercadeo. Recuperado de: https://repository.usta.edu.co/bitstream/handle/11634/12525/2018gustavovenegas.pdf?sequence=1 .

El grafico del dendograma que se observa en la figura 3-7, ilustra todos los grupos existentes después de agruparlos por características y patrones de similitud que se encuentren. Definir cuantos grupos se desean generar o cuantos tipos de planetas o etiquetas se poseen con un corte, aproximara más las predicciones realizadas a los resultados esperados o deseados del modelo predictivo.

En la figura 3-8 se encuentra el diagrama de flujo del algoritmo de clustering jerarquico.se inicia el algoritmo con la creación de los primeros grupos por el método ward y a partir de los datos obtenidos se genera el primer dendograma. A continuación se realiza un corte en el dendograma en el punto t=45 para crear los 4 grupos o etiquetas requeridas. Por último, se imprime la columna de grupos generada a partir del modelo implementado, para definir si la clasificación realizada es acertada.

Figura 3-8: Diagrama de flujo del algoritmo clustering jerarquico.



4. Capítulo 4

4.1 Resultados y Análisis de Resultados.

Una vez aplicadas las técnicas y algoritmos de MLen la base de datos previamente procesada, se procedió a realizar la tabulación y graficación de los resultados obtenidos con cada uno de los algoritmos, para posteriormente analizar los resultados y con esta información recolectada, generar las conclusiones a las cuales llega la investigación y determinar cuál de todos los algoritmos implementados es mejor al estimar patrones no identificados previamente.

En la figura 4-1 se observa la tabla que contiene los valores de precisión y exactitud obtenidos al implementar los algoritmos de regresión logística, k vecino más cercano y árbol de decisiones de las técnicas de aprendizaje de Machine Lerning supervisado.

Inicialmente se implementó el algoritmo con los datos que se poseen referentes a la tierra y sin pasarlos a escala logarítmica, considerando que la correlación existente entre los atributos empleados es menor en esta instancia, se observa que los resultados de precisión y exactitud obtenidos son menores a los generados al implementar los mismos algoritmos con los datos en escala logarítmica como se puede apreciar en la figura 4-5

.

Figura 4-1: Resultados de las métricas de precisión y exactitud de los algoritmos de aprendizaje supervisado.

Index	Nombre Algoritmo Supervisado	Precision	Exactitud
0	Regresion Logistica	0.858663	0.858663
1	KVecinosMasCercanos	0.8769	0.8769
2	Arbol Decisiones	0.656535	0.656535

Nombre de la fuente: Spyder(Python 3.8).

La matriz de confusión obtenida al implementar el algoritmo de regresión logística que se observa en la figura 4-2, nos indica con más detalle el número de aciertos y el número de errores generados con la predicción realizada. Dicha matriz posee 4 filas en representación a las 4 etiquetas o tipos de planetas que se definieron inicialmente. Los valores que se observan en la diagonal principal son los valores estimados de forma correcta por el modelo predictivo generado.

La precisión se calcula empleando la ecuación 2. Donde el valor de los verdaderos positivos (TP) es la suma de la diagonal principal igual a 565 y los falsos positivos (FP) es igual a la suma de los planetas clasificados como si pertenecieran a un tipo pero no lo son, igual a 93, como se muestra en la ecuación 3 mostrada a continuación:

$$Precision = TP/(TP + FP) = {}^{565}/_{(565 + 93)} = 0.8586 \quad (3)$$

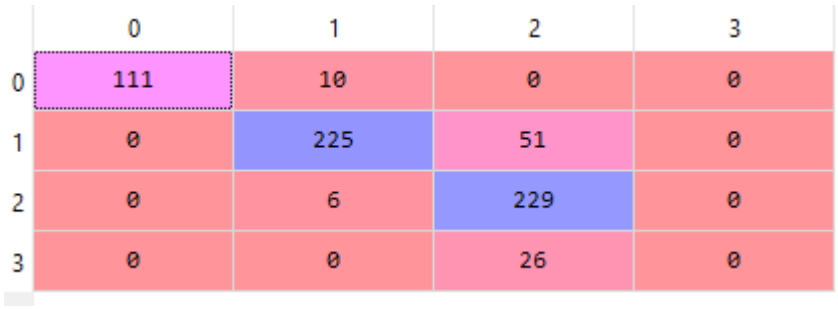
La exactitud se calcula empleando la ecuación 1.Donde el valor de los verdaderos positivos (TP) continua siendo 565 y los falsos positivos (TN) es igual a todos los valores marcados como verdaderos y que no pertenecen a la clase designada, igual a 0, como se muestra en la ecuación 4 mostrada a continuación:

$$Exactitud = {}^{TP + TN}/_{Total} = {}^{565}/_{658} = 0.8586 \quad (4)$$

El valor de precisión y exactitud que se obtuvo en la ecuación 3 y 4 respectivamente, corresponden al resultado generado con el algoritmo de regresión logística como se

aprecia en la figura 4-1, calculados a partir de los valores resultantes obtenidos en la matriz de confusión de la figura 4-2.

Figura 4-2: Matriz de confusión del algoritmo regresión logística.



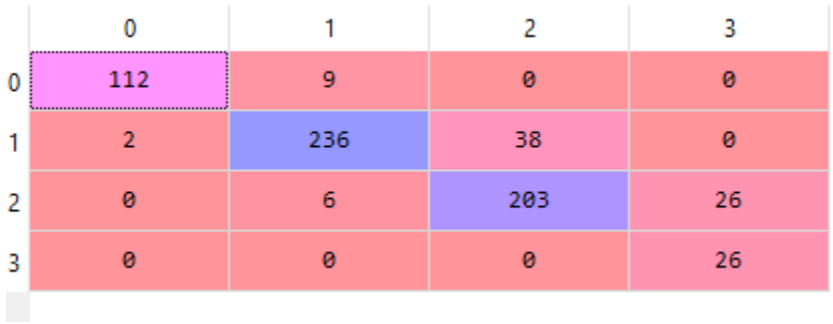
	0	1	2	3
0	111	10	0	0
1	0	225	51	0
2	0	6	229	0
3	0	0	26	0

Nombre de la fuente: Spyder(Python 3.8).

En la figura 4-3 se aprecia la matriz de confusión al implementar el algoritmo de k vecinos más cercanos, en esta matriz se observa que la suma de su diagonal principal o los valores estimados correctamente es mayor a la presentada en la matriz de confusión del algoritmo de regresión logística en la figura 4-2.

Debido a lo mencionado con anterioridad se puede inferir que los valores obtenidos al calcular la precisión y exactitud del modelo de k vecinos va ser mayor a la obtenida con el algoritmo de regresión logística como se muestra en la figura 4-1.

Figura 4-3: Matriz de confusión del algoritmo k vecinos más cercanos.



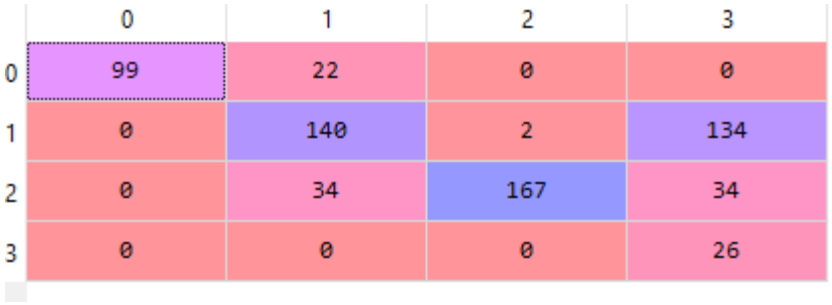
	0	1	2	3
0	112	9	0	0
1	2	236	38	0
2	0	6	203	26
3	0	0	0	26

Nombre de la fuente: Spyder(Python 3.8).

En la figura 4.4 se aprecia la matriz de confusión al implementar el algoritmo de árbol de decisión. En esta matriz se observa que la suma de su diagonal principal de los valores estimados correctamente por el modelo es mucho menor con respecto a las matrices de

confusión de los algoritmos de regresión logística y k vecinos analizadas con anterioridad. Debido a esto se puede inferir que los valores obtenidos al calcular la precisión y exactitud del modelo de árbol de decisión van a ser mucho menores que los demás algoritmos implementados como se muestra en la figura 4-1.

Figura 4-4: Matriz de confusión del algoritmo Árbol de decisión.



	0	1	2	3
0	99	22	0	0
1	0	140	2	134
2	0	34	167	34
3	0	0	0	26

Nombre de la fuente: Spyder(Python 3.8).

Recordando que la correlación es mayor con los datos en escala logarítmica, a la hora de implementar los mismos algoritmos mencionados anteriormente y efectuar el cálculo de las mismas métricas de evaluación del modelo predictivo con dichos datos, se esperaba obtener un resultado mayor con respecto a los valores de precisión y exactitud estudiados. Este comportamiento se observa en la tabla de la figura 4-5 que contiene las métricas de evaluación de los algoritmos de aprendizaje supervisado con los datos en escala logarítmica.

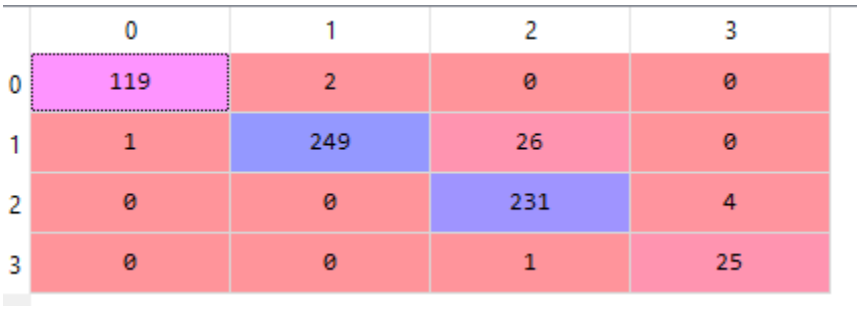
Figura 4-5: Resultados de las métricas de precisión y exactitud de los algoritmos de aprendizaje supervisado con los datos en escala logarítmica.

Index	Nombre Algoritmo Supervisado	Precision	Exactitud
0	Regresion Logistica	0.919453	0.919453
1	KVecinosMasCercanos	0.93465	0.93465
2	Arbol Decisiones	0.948328	0.948328

Nombre de la fuente: Spyder(Python 3.8).

Los resultados obtenidos con los datos logarítmicos son más precisos y exactos a los obtenidos con los datos iniciales.

Figura 4-6: Matriz de confusión del algoritmo Árbol de decisión con los datos en escala logarítmica.

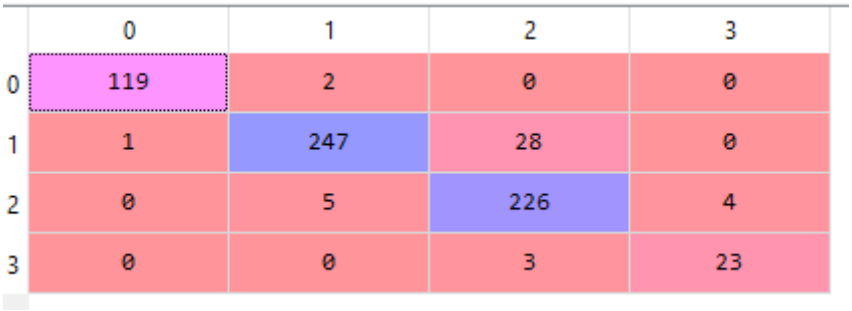


	0	1	2	3
0	119	2	0	0
1	1	249	26	0
2	0	0	231	4
3	0	0	1	25

Nombre de la fuente: Spyder(Python 3.8).

A diferencia de los resultados obtenidos con los datos iniciales, al implementar los datos en escala logarítmica se obtiene que el modelo predictivo más acertado es el árbol de decisiones, el cual presenta una matriz de confusión que se muestra en la figura 4-6, con una suma de los valores de su diagonal principal o valores predichos correctamente mayor con respecto a los resultados de la suma de las diagonales principales de los demás algoritmos.

Figura 4-7: Matriz de confusión del algoritmo k vecinos más cercanos con los datos en escala logarítmica.



	0	1	2	3
0	119	2	0	0
1	1	247	28	0
2	0	5	226	4
3	0	0	3	23

Nombre de la fuente: Spyder(Python 3.8).

La matriz de confusión de la figura 4.7 del algoritmo de k vecinos fue el más acertado con los datos iniciales con respecto a los demás algoritmos, ahora con los datos logarítmicos aumento su exactitud y predicción pero fue superado por el algoritmo de árbol de decisión.

Figura 4-8: Matriz de confusión del algoritmo regresión logística con los datos en escala logarítmica.

	0	1	2	3
0	113	8	0	0
1	1	243	32	0
2	0	6	227	2
3	0	0	4	22

Nombre de la fuente: Spyder(Python 3.8).

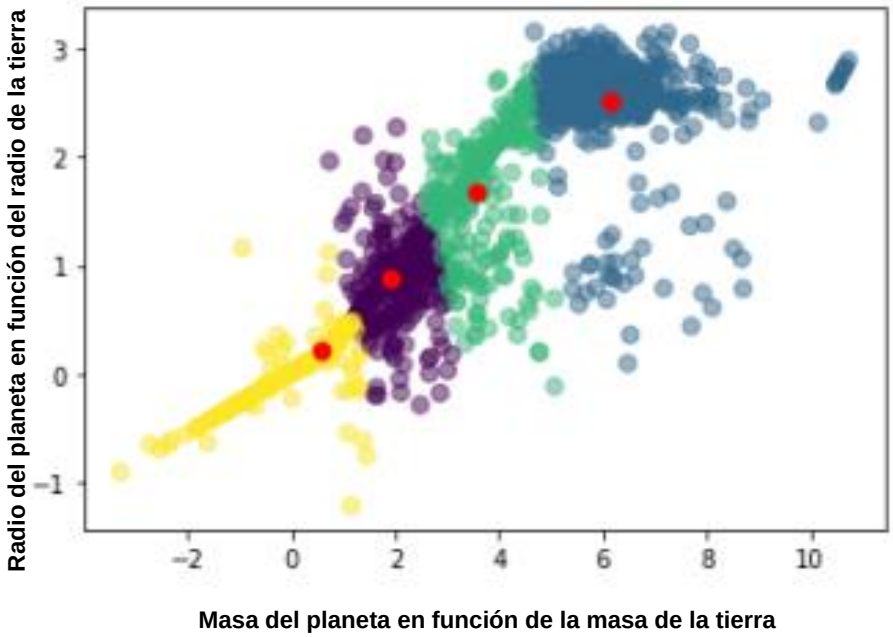
Por último, el modelo predictivo del algoritmo de regresión logística fue el menos acertado entre los tres algoritmos de aprendizaje supervisados implementados como se observa en la figura 4-5.Este algoritmo al igual que los demás, aumentaron su precisión y exactitud, incrementando el número de aciertos obtenidos en su matriz de confusión figura 4-8 con los datos en escala logarítmica, obteniendo un muy buen porcentaje de acierto.

A continuación se encuentran los resultados obtenidos al implementar los algoritmos de aprendizaje no supervisado.

En el caso del algoritmo de k means se generó una gráfica que se observa en la figura 4-9 en donde se puede observar los 4 centroides definidos y a partir de ellos se agrupan los datos en 4 grupos o se clasifican los exoplanetas en los 4 tipos de planetas.

Inicialmente se puede observar que la predicción generada a partir de este algoritmo es muy aproximada a la clasificación que se observa en la figura 2-12 donde se observa la gráfica de dispersión del radio del planeta con respecto a la masa del planeta en función de la tierra.

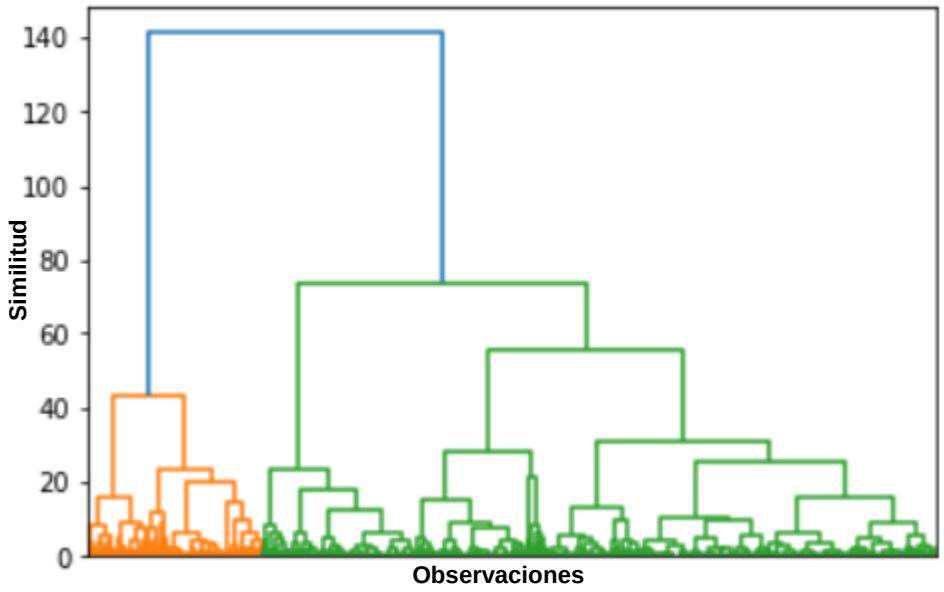
Figura 4-9: Grafica resultante del Algoritmo de k-medias o k-means. masa del planeta vs el radio del planeta.



Nombre de la fuente: Spyder(Python 3.8).

Al implementar el algoritmo de clustering jerarquico y generar el dendograma de la figura 4-10 con los datos iniciales sin indicarle el número de grupos que se desea generar, se observa que inicialmente se crean multiples grupos con respecto a la similitud encontrada entre los datos de la base de exoplanetas.

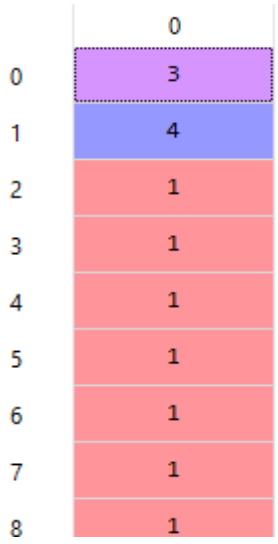
Figura 4-10 Grafica Dendograma del algoritmo Clustering Jerarquico.



Nombre de la fuente: Spyder(Python 3.8).

Al generar el corte en el dendograma para indicarle el número de grupos o clusters que se desean encontrar, se genera una columna de clustering como se observa en la figura 4-11 donde se clasifica cada planeta en los 4 grupos definidos por el corte realizado.

Figura 4-11: Columna generada de los Clustering creados a partir del Dendograma obtenido.



	0
0	3
1	4
2	1
3	1
4	1
5	1
6	1
7	1
8	1

Nombre de la fuente: Spyder(Python 3.8).

Por último se incluyó la columna nueva de clusters o grupos a la tabla donde se tiene toda la información procesada que se posee sobre los exoplanetas observados para obtener una sola tabla como se observa en la figura 4-12 y divisar de una manera más clara las diversas clasificaciones o predicciones realizadas por el modelo y definir qué tan acertado fue el modelo realizando generalizaciones o definiendo a cuál de los 4 tipos de planetas conocidos pertenece cada exoplaneta dentro de la base de datos.

Figura 4-12: Tabla final con la columna de clustering creados a partir del dendograma obtenido.

Index	Planet_Name	Planet_Radius(Earth)	Orbital_Period(Days)	Planet_Mass_Earth	Planet_Type	Clustering Jerarquico
16	2MASS J21402931+1625183 A b	2.33311	8.90062	8.80419	Gas Giant	3
51	BD+20 594 b	0.947014	3.73015	3.10226	Neptune-like	4
68	CoRoT-1 b	2.81541	0.411418	5.79161	Gas Giant	1
69	CoRoT-10 b	2.38601	2.58329	6.77365	Gas Giant	1
70	CoRoT-11 b	2.77446	1.09672	6.60792	Gas Giant	1
71	CoRoT-12 b	2.7813	1.03958	5.6754	Gas Giant	1
72	CoRoT-13 b	2.29455	1.39505	6.03055	Gas Giant	1
73	CoRoT-14 b	2.50307	0.413526	7.7902	Gas Giant	1
74	CoRoT-16 b	2.57338	1.67752	5.13656	Gas Giant	1
75	CoRoT-17 b	2.43624	1.32657	6.64994	Gas Giant	1
76	CoRoT-18 b	2.68649	0.64189	7.00621	Gas Giant	1
77	CoRoT-19 b	2.67139	1.36024	5.86641	Gas Giant	1
78	CoRoT-2 b	2.79923	0.555604	7.00621	Gas Giant	1
81	CoRoT-21 b	2.6791	1.00237	6.57742	Gas Giant	1
82	CoRoT-22 b	1.58515	2.27788	2.50144	Neptune-like	4
83	CoRoT-23 b	2.46555	1.28959	6.79167	Gas Giant	1
84	CoRoT-24 b	1.30833	1.63186	1.74047	Neptune-like	4
85	CoRoT-24 c	1.60944	2.46462	3.3322	Neptune-like	1
86	CoRoT-25 b	2.49403	1.58118	4.45272	Gas Giant	1
87	CoRoT-26 b	2.64759	1.43621	5.10812	Gas Giant	1
88	CoRoT-27 b	2.42365	1.27405	8.1029	Gas Giant	1

Nombre de la fuente: Spyder(Python 3.8).

Con los algoritmos de aprendizaje no supervisado no existe seguridad sobre el comportamiento del clasificador. Es un método menos preciso y fiable ya que no se puede obtener información precisa con respecto a la clasificación de los datos y la salida, a diferencia del aprendizaje supervisado cuyos datos están etiquetados y se conocen las respuestas.

5. Conclusiones y recomendaciones

5.1 Conclusiones

Teniendo en cuenta los objetivos planteados inicialmente en este trabajo, los resultados expuestos son la clara representación de que el proceso de implementar las técnicas y algoritmos de aprendizaje supervisado y no supervisado de ML en el análisis de datos astronómicos sobre exoplanetas obtenidos a través de diversos telescopios terrestres, espaciales y diferentes misiones realizadas, entre ellos: Tess, Kepler, K2, KELT y Ukirt, son muy acertadas a la hora de generar modelos predictivos que permitan realizar una correcta clasificación de los planetas extrasolares observados. Además, los mencionados resultados son perfectos para realizar un análisis de los algoritmos implementados y determinar cuál de todos es mejor al estimar patrones no identificados previamente.

Estos resultados son expuestos empleando diversas gráficas y tablas que permiten observar y comparar más claramente la información obtenida a través de la implementación de los mencionados algoritmos.

Se concluyó que los algoritmos de aprendizaje no supervisado no permiten tener una seguridad sobre el comportamiento del clasificador, debido a que no se posee una etiqueta inicial que permita la clasificación de los datos para el entrenamiento del modelo ingresando las respuestas esperadas. Sin embargo, el hecho de conocer el número de etiquetas existentes, permite generar exactamente los grupos requeridos y obtener un mejor resultado. En conclusión, el aprendizaje no supervisado es un método menos preciso y fiable ya que no se puede obtener información precisa con respecto a la clasificación de los datos y la salida, a diferencia del aprendizaje supervisado cuyos datos están etiquetados y se conocen las respuestas.

Al implementar los algoritmos de aprendizaje supervisa se obtuvieron mejores resultados y todos los algoritmos arrojaron una precisión y una exactitud bastante alta. El algoritmo que destaco al entregar los mejores resultados registrados fue el algoritmo de árbol de decisión con un 94,8328% de precisión y exactitud. Implementar los datos en escala logarítmica y no los iniciales mejoro notoriamente los porcentajes de precisión y exactitud obtenidos inicialmente.

5.2 Recomendaciones

El objetivo a largo plazo es poder seguir implementando estas técnicas y algoritmo de ML incorporando los datos nuevos que se agreguen a la base de datos. Ya que el universo es tan extenso y día a día se realizan nuevas observaciones desde cada observatorio y empleando diversos métodos, se espera que la base de datos crezca exponencialmente.

Definir más atributos relevantes que mejoren los resultados, al igual que realizar una mejor división de los datos para una remarcada separación de los grupos existentes, podría generar que al implementar los algoritmos de aprendizaje no supervisado se obtenga una exactitud y precisión más alta en los resultados obtenidos al hacer más remarcada la separación y con la disminución de los datos sobrelapados, se lograría generar un agrupamiento más efectivo de la información.

Bibliografía

Liga Iberoamericana de Astronomía. (febrero de 2014). *Sección de Exoplanetas – Planetas extrasolares / LIADA Liga Iberoamericana de Astronomía*. Obtenido de https://exoplanetasliada.wordpress.com/metodos-de-deteccion/

Bagnato, J. I. (10 de Julio de 2018). *Aprendamos Machine Learning*. Obtenido de https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/#:~:text=%C2%BFQu%C3%A9%20es%20el%20algoritmo%20k,de%20datos%20que%20le%20rodean.

BSA data study. (2015). Obtenido de https://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_es.pdf

Caparrini, F. S. (14 de Diciembre de 2020). *Fernando Sancho Caparrini*. Obtenido de http://www.cs.us.es/~fsancho/?e=77

CASTILLO, N. M. (2018). *DEEP LEARNING PARA IDENTIFICACIÓN DE NÚCLEOS ACTIVOS DE GALAXIAS POR VARIABILIDAD.* Santiago de Chile .

Chauhan, N. S. (2 de Septiembre de 2020). *DataSource.AI*. Obtenido de DATASOURCE.AI: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Dera, D., & Bhavsar, P. (2017). *Aprendizaje automático en análisis de datos de transporte.* Obtenido de sciencedirect: https://www.sciencedirect.com/topics/psychology/machine-learning

Exoplanet Exploration. (2 de mayo de 2021). Obtenido de https://exoplanets.nasa.gov/discovery/exoplanet-catalog/

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

Fuentes, D. P. (2019). *Aplicación de algoritmos de Machine Learning para la predicción del beneficio por cliente a partir de métricas de Google Analytics.* Valladolid.

Heras, J. M. (20 de Septiembre de 2020). *Iartificial*. Obtenido de https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/

HOJAS, I. M. (s.f.). *Stat Developer*. Obtenido de https://www.statdeveloper.com/agrupacion-en-cluster-jerarquica-en-python/

KELT. (2021). Obtenido de https://keltsurvey.org/about

LucidChart. (2021). Obtenido de https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision

Machine Learning,una pieza clave en la transformación de los modelos de negocio. (2018). Management solutions.

Management Solutions. (2018). Obtenido de https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Mercado, J. A. (19 de septiembre de 2017). *big data dummy*. Obtenido de https://bigdatadummy.wordpress.com/2017/01/15/metodos-de-regresion-clasificacion-y-clustering/

NASA. (2021). *NASA Exoplanet Archive*. Obtenido de https://exoplanetarchive.ipac.caltech.edu/

Osorio, J. K. (2019). *METODOLOG´IA DE CLASIFICACION DE DATOS DESBALANCEADOS BASADO EN METODOS DE SUBMUESTREO.* Pereira.

Raschka, S., & Mirjalili, V. (2019). *Aprendizaje automático de Python: tercera edición.* Packt. Obtenido de https://exoplanetarchive.ipac.caltech.edu/

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

RUIZ, J. Z. (2018). *Comparativas de Analisis de Algoritmos de Aprendisaje Automatico para la Prediccion del Tipo Predominante de Cubierta Arborea.* Madrid.

Sandoval, L. J. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. *ITCA*.

SINCLAIR, R. M. (2016). *CARACTERIZACIÓN DE EXOPLANETAS MEDIANTE TÉCNICAS DE PROCESAMIENTO DE SEÑALES Y MÉTODO DE LAS VELOCIDADES RADIALES.*

Soporte de miniTab18. (2019). Obtenido de https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/

Venegas, G. A., & Pineda Rios, W. (2017). *Metodo de Cluster Jerarquicos para Datos Funcionales Multivariados: Una Aplicacion en Mercadeo.*