ISSN 1870-4069

Técnicas de aprendizaje automático para agrupar planetas similares a la Tierra

Marcos Macías-Juárez, Edgar Moyotl-Hernández

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias Físico Matemáticas,
México

marcos.macias@alumno.buap.mx, emoyotl@fcfm.buap.mx

Resumen. Este trabajo presenta un modelo de aprendizaje automático, con técnicas de agrupamiento y clasificación, que puede utilizarse para formar un grupo de exoplanetas con características similares a las de la Tierra. Para la parte de clasificación, se utilizó k-nn y para la parte del agrupamiento, se uso k-means++, DBSCAN, agglomerative clustering y divisive clustering. Aplicando el modelo a un conjunto de datos con información de más de 5000 exoplanetas, descubiertos por la NASA, se identificó que el mejor modelo de aprendizaje se consigue con agglomerative clustering o divisive clustering y k-nn, juntos logran encontrar un grupo de 66 exoplanetas con características comparables a las del planeta Tierra. Los resultados obtenidos muestran que las técnicas de aprendizaje automático pueden ser aplicadas en diversas áreas.

Palabras clave: Aprendizaje automático, agrupamiento, clasificación, astronomía, exoplanetas.

Machine Learning Techniques to Group Earth–like Planets

Abstract. This work presents a machine learning model that uses clustering and classification techniques which can be used to form a group of exoplanets with characteristics similar to those of the Earth. For the classification part, we used k-nn and for the clustering part, we used k-means++, DBSCAN, agglomerative clustering and divisive clustering. Applying the model on a dataset with information on more than 5000 exoplanets discovered by NASA, it was identified that the best learning model is achieved with agglomerative clustering or divisive clustering and k-nn, together they achieve to find a cluster of 66 exoplanets with characteristics comparable to those of the planet Earth. The results obtained show that machine learning methods can be applied in various areas.

Keywords: Machine learning, clustering, classification, astronomy, exoplanets.

1. Introducción

Una de las líneas de estudio en astrofísica se encarga de la búsqueda de exoplanetas y del estudio de los mismos. Un exoplaneta es un planeta que se encuentra fuera del sistema solar, esto significa que orbita alrededor de una estrella central distinta al Sol. Los exoplanetas se estudian con el propósito de detectar planetas con características similares a las de la Tierra, esto es de suma importancia dado que, analizando la composición de estos planetas y teniendo en cuenta factores como la distancia entre este y su estrella anfitriona, es posible deducir teóricamente si podrían albergar vida.

La misión Kepler de la NASA revolucionó el campo de la exploración de exoplanetas y ha permitido a los investigadores recopilar datos sobre objetos extrasolares de interés. Para manejar la enorme cantidad de datos recolectados, es importante desarrollar métodos de análisis de datos que sean capaces de identificar y confirmar que un objeto dado sea un exoplaneta y más aún, métodos para clasificar exoplanetas (según sus propiedades físicas) en categorías apropiadas de habitabilidad para dar una idea de qué tan similar o diferente es un exoplaneta a la Tierra.

El aprendizaje automático constituye esa herramienta capaz de tratar con enormes cantidades de información, debido a esto las técnicas y algoritmos de aprendizaje automático se han vuelto muy importantes en la astronomía y en muchos otros campos, siendo empleados en múltiples tareas. Este trabajo se centra en la aplicación de técnicas de aprendizaje automático supervisado y no supervisado, en primer lugar, para entrenar un modelo capaz de agrupar exoplanetas con características similares entre ellos; y en segundo lugar, para clasificar a los planetas del sistema solar, en particular el planeta Tierra, en alguno de estos grupos.

Con este modelo, se busca un grupo de planetas con características parecidas a las de la Tierra y así, posiblemente, encontrar algún planeta en el cuál sea posible albergar la vida que alberga la Tierra. En este estudio, se utilizó un conjunto de datos de la NASA que contiene características de exoplanetas confirmados. Además, para la parte de aprendizaje supervisado, se utilizó el algoritmo k-nn y para la parte de aprendizaje no supervisado, se usaron los métodos k-means++, DBSCAN, agglomerative clustering y divisive clustering.

El artículo está organizado de la siguiente manera. En la sección 2 se exponen los trabajos relacionados. La sección 3 presenta la propuesta de solución donde se explica la metodología que se siguió. En la sección 4 se describen las pruebas con el modelo propuesto. En la sección 5 se analizan y discuten los resultados obtenidos de los experimentos. Finalmente, la sección 6 muestra las conclusiones y el trabajo futuro.

2. Trabajo relacionado

Existen diversos estudios que exploran la eficacia del aprendizaje automático para la detección de exoplanetas y para clasificar exoplanetas en clases de habitabilidad, basándose en las características físicas de los propios exoplanetas y sus estrellas madre o anfitrión. En [3] el objetivo del estudio fue comparar la eficiencia de los métodos k-nn, logistic regression y decision trees con respecto a la detección de exoplanetas.

Tabla 1. Características de cada exoplaneta.

Característica	Descripción
name	Nombre del planeta dado por la NASA.
distance	Distancia del planeta a la tierra en años luz.
stellar_magnitude	Brillo del planeta.
planet_type	Tipo de planeta derivado de los del sistema solar.
discovery_year	Año en que se descubrió el planeta.
mass_multiplier	Multiplicidad de masa del planeta con el planeta de 'mass_wrt'.
mass_wrt	Masa comparada a la de algún planeta del sistema solar.
radius_multiplier	Multiplicidad de radio del planeta con el planeta de 'radio_wrt'.
radius_wrt	Radio comparado al de algún planeta del sistema solar.
orbital_radius	Radio orbital de los planetas alrededor de su Sol (en AU).
orbital_period	Años que tardan en completar 1 órbita de su estrella anfitrión.
eccentricity	Indica qué tan circular es la trayectoria orbital del planeta.
detection_method	Método utilizado por la NASA para encontrar ese planeta.

Utilizaron estos algoritmos de aprendizaje supervisado para analizar un conjunto de datos con muestras tanto reales como sintéticas y entrenar un modelo capaz de predecir con precisión si un objeto astrofísico es un exoplaneta o no. Los resultados mostraron que k-nn, al ser un clasificador no lineal, fue el más eficiente en clasificar correctamente los exoplanetas, logrando una precisión del 98.22 %.

Por su parte, en [4] llevaron a cabo aprendizaje supervisado y no supervisado en dos conjuntos de datos recopilados por la NASA, el conjunto de datos Kepler y un conjunto de datos de exoplanetas confirmados, el NASA Exoplanet Archive. Por un lado, al conjunto de datos Kepler lo usaron para predecir la existencia de candidatos a exoplanetas como tarea de clasificación, utilizando los algoritmos: decision trees, random forest, naïve Bayes y redes neuronales.

Como resultado, los métodos obtuvieron precisiones del 99.06 %, 92.11 %, 88.50 % y 99.79 %, respectivamente. Por otro lado, el conjunto de datos de exoplanetas confirmados lo usaron para encontrar exoplanetas habitables, como tarea de agrupamiento. Para esto, dividieron los exoplanetas confirmados en diferentes grupos utilizando el algoritmo k-means. Antes de agrupar los exoplanetas agregaron las características de la Tierra al conjunto de datos. Luego, dividieron todos los planetas en 100 clusters y consideraron que los exoplanetas que tienen más probabilidades de ser habitables se encontraban en el grupo que contenía a la Tierra.

Esto significa que encontraron un grupo de 100 exoplanetas con caracterısticas parecidas a las de la Tierra. En contraste con los trabajos citados, este trabajo, se centra en la aplicación de la clasificación y el agrupamiento para la detección de exoplanetas con características similares a las de la Tierra. Primero, agrupa exoplanetas con características similares entre ellos; y segundo, clasifica a los planetas del sistema solar, en alguno de estos grupos.

Tabla 2. Características de los planetas del sistema solar.

Planeta	mass_multiplier	radius_multiplier	orbital_radius	eccentricity
Mercurio	0.055	0.38	0.38	0.2000
Venus	1.000	1.00	0.99	0.0170
Tierra	0.820	0.95	0.72	0.0068
Marte	0.500	0.53	1.50	0.0930
Júpiter	318.000	11.00	5.20	0.0480
Saturno	95.180	9.40	9.50	0.0566
Urano	0.140	4.00	19.10	0.0460
Neptuno	17.000	3.80	30.06	0.0097

Los ocho planetas del sistema solar, en orden de cercanía al Sol, son: Mercurio, Venus, Tierra, Marte, Júpiter, Saturno, Urano y Neptuno. La clasificación y el agrupamiento son técnicas para encontrar patrones usadas en el aprendizaje automático (machine learning, en Inglés). La tarea de clasificación se enmarca en el aprendizaje supervisado, se enfoca en la creación de modelos a partir de un conjunto de datos etiquetados (es decir, datos para los que ya se conoce la respuesta correcta) y permite predecir a que clase pertenece un objeto nuevo.

En cambio, la técnica de agrupamiento o clustering pertenece al aprendizaje no supervisado y su objetivo es crear modelos, a partir de un conjunto de datos no etiquetados, que encuentren grupos de objetos similares de forma que los elementos del mismo grupo estén más relacionados entre ellos que aquellos elementos de diferentes grupos [7], estos grupos se conocen como clústeres o clusters.

3. Metodología

El modelo de aprendizaje que se propone consta de dos partes, algoritmos de aprendizaje supervisado para clasificación y algoritmos de aprendizaje no supervisado para clustering. Para la parte de clasificación, se utilizó k-nn y para la parte del agrupamiento, se uso k-means++, DBSCAN, agglomerative clustering y divisive clustering. Así mismo, las medidas de calidad para determinar el rendimiento del modelo propuesto fueron el coeficiente de silueta y el índice de Davies Bouldin.

3.1. Método K-NN

El algoritmo de k vecinos más cercanos, también conocido como k-nn (del inglés k-nearest neighbours), es un clasificador de aprendizaje supervisado no paramétrico y no lineal, que utiliza la proximidad entre puntos de datos para hacer predicciones sobre la clasificación de un punto individual. Al final, el nuevo objeto pertenece a la categoría con mayoría de votos entre su k vecinos más cercanos. La “k” en k-nn representa el número de vecinos más cercanos considerados [7].

Algorithm 1: Modelo de aprendizaje propuesto

Input: Conjunto de exoplanetas X, conjunto de planetas del sistema solar P, número máximo de iteraciones $MAXITE$.
Output: Un grupo de planetas similares $C_t \subset X$.

for $i \leftarrow 1$ **to** $MAXITE$ **do**
  $k \leftarrow$ card(P) ; // Cardinalidad del conjunto P
  $C \leftarrow$ agrupar(X, k);
  for $j \leftarrow 1$ **to** k **do**
    $y_j \leftarrow$ clasificar(p_j, C);
    if $p_j =$ *Tierra* **then**
      $t \leftarrow y_j$; // Etiqueta del grupo de la Tierra
    end
  end
  $S \leftarrow P \cap C_t$; // Planetas en el grupo de la Tierra
  if *card*$(S) = 1$ **then**
    return C_t;
  end
  $X \leftarrow C_t$;
  $P \leftarrow S$;
end

3.2. Métodos k-means y k-means++

k-means es un algoritmo de aprendizaje no supervisado que agrupa objetos en k grupos basándose en prototipos.

El agrupamiento basado en prototipos significa que cada grupo se representa por un centroide o por un medoide [7]. De acuerdo con la información presentada en [2], k-means es el algoritmo de agrupación particional más utilizado. El método k-means es muy bueno para identificar grupos con una forma esférica pero uno de sus inconvenientes es que se tiene que especificar, con antelación, la cantidad k de grupos a generar [7]. Por su parte, el algoritmo k-means++ selecciona cuidadosamente los centroides iniciales para la agrupación de k-means y luego se realiza la agrupación con k-means clásico utilizando estos centroides [1].

3.3. Método DBSCAN

Entre los métodos de agrupamiento, los algoritmos basados en densidad enfocan el problema de dividir el conjunto de datos en grupos teniendo en cuenta la distribución de densidad de los puntos [6]. DBSCAN (Density Based Spatial Clustering of Applications with Noise) es el primer algoritmo de agrupamiento basado en densidad [5]. Tiene dos parámetros principales que son el radio de la vecindad (eps) que es la distancia máxima entre dos puntos para poder ser considerados pertenecientes al mismo vecindario, y el número mínimo de puntos (minpts) en un vecindario para que un punto pueda ser considerado de alta densidad. DBSCAN puede resolver problemas en los cuales k-means puede fallar puesto que funciona muy bien con formas complejas que no tienen que ser esféricas, además, identifica los valores atípicos o ruido en los datos y, no necesita que se defina de antemano el número de clústeres [6].

Tabla 3. Cantidad de planetas por cluster (primera iteración).

	Algoritmos de agrupamiento			
Cluster	**k-means++**	**DBSCAN**	**Agglomerative**	**Divisive**
0	**4526**	1717	245	2
1	1	**3002**	44	7
2	2	7	7	**3909**
3	1	12	**3909**	498
4	5	6	498	245
5	224	9	2	59
6	5	6	59	44
7	1	6	1	1

3.4. Método Agglomerative clustering

Los métodos jerárquicos construyen una jerarquía de clústeres en el conjunto de datos. Esta jerarquía se representa como un árbol (o dendrograma). La raíz del árbol es el único grupo que contiene a todos los puntos de datos, siendo las hojas los grupos con un sólo dato. Hay dos tipos de algoritmos de agrupamiento jerárquico con enfoques inversos: el algoritmo aglomerativo (agglomerative clustering) y el algoritmo divisivo (divisive clustering).

El algoritmo jerárquico aglomerativo, es una estrategia ascendente que inicia considerando cada punto de datos como un grupo individual y luego va fusionando pares de grupos similares progresivamente hasta que todos los datos pertenezcan a un único grupo [7].

3.5. Método divisive clustering

El algoritmo jerárquico divisivo es inverso al aglomerativo, sigue un enfoque de arriba hacia abajo, comienza con un sólo grupo que contiene todos los puntos de datos y lo va dividiendo iterativamente en grupos más pequeños hasta obtener un grupo para cada dato [7]. El método aglomerativo, generalmente, resulta ser más sencillo de implementar que el método divisivo porque existe un único modo de unir dos grupos, mientras que existen muchas maneras de separar un conjunto de puntos en dos grupos, lo cual es muy costoso en términos computacionales. Por otro lado, una de las ventajas de este tipo de algoritmos sobre k-means es que no requiere que los datos se representen en espacios vectoriales, ya que solo requiere de una medida de distancia entre puntos. Tampoco requiere especificar por anticipado el número de grupos. Sin embargo, pueden producir grupos de tamaños desiguales.

3.6. Métricas de calidad

Las técnicas de aprendizaje no supervisado, a diferencia de las supervisadas, se entrenan con conjuntos de datos sin etiquetas, por lo que, cuantificar la calidad de los resultados proporcionados por un algoritmo de agrupamiento no es un problema fácil.

Tabla 4. Resultados de agglomerative clustering por iteración.

	Agglomerative clustering						
	Iteración 1		**Iteración 2**	**Iteración 3**		**Iteración 4**	
Cluster	**EA**	**PC**	**PC**	**EA**	**PC**	**EA**	**PC**
0	245	0	1	763	0	85	2
1	44	0	0	587	0	86	1
2	7	0	0	879	1	79	0
3	**3909**	**8**	**5**	269	0	**66**	**1**
4	498	0	0	**316**	**4**	-	-
5	2	0	1	-	-	-	-
6	59	0	0	-	-	-	-
7	1	0	1	-	-	-	-
S	0.8286		0.6304	0.4967		0.5012	
I_DB	0.4257		0.5409	0.5519		0.6080	

En este sentido, el coeficiente de silueta y el índice de Davies Bouldin son medidas de calidad de los grupos resultantes de un agrupamiento y pueden aplicarse a distintos algoritmos. Las dos métricas calculan la cohesión (distancia promedio de un punto al resto de puntos dentro del mismo grupo) y la separación (distancia promedio de un punto a todos los puntos en el grupo más cercano) [7].

El coeficiente de silueta S se calcula como la diferencia entre la cohesión y la separación del grupo dividida por el valor más grande de los dos; S puede tomar valores entre -1 y 1, siendo 1 el valor ideal. En cambio, el índice de Davies Bouldin I_{DB} se calcula como la relación promedio entre la cohesión y la separación; los valores de I_{DB} más próximos a 0 son mejores.

4. Experimentación

El objetivo del presente estudio es comparar la eficiencia de varios algoritmos de aprendizaje automático respecto a la agrupación de exoplanetas.

4.1. Conjunto de datos

Para los experimentos de aprendizaje no supervisado se utilizó el NASA Exoplanet Archive, este conjunto de datos contiene información sobre los exoplanetas descubiertos por la NASA[1] en diversas misiones espaciales, observatorios terrestres y otras fuentes[2]. Este conjunto de datos de planetas (ya confirmados como exoplanetas) incluye información como el nombre del planeta, masa, radio, distancia desde su estrella

[1] https://www.nasa.gov

[2] Los datos están accesibles en la plataforma Kaggle en el siguiente enlace: https://www.kaggle.com/datasets/adityamishraml/nasaexoplanets

Tabla 5. Resultados de k-means por iteración.

	k-means++										
	Iteración 1		**Iteración 2**		**Iteración 3**		**Iteración 4**		**Iteración 5**		**Iteración 6**
Cluster	EA	PC	EA	PC	EA	PC	EA	PC	EA	PC	PC
0	**4526**	**8**	**4185**	**8**	137	0	1000	1	**1447**	**4**	**4**
1	1	0	13	0	41	0	21	1	1	0	0
2	2	0	1	0	130	0	1	0	5	0	0
3	1	0	122	0	**3272**	**6**	144	1	621	1	0
4	5	0	2	0	154	1	**2079**	**5**	5	0	-
5	224	0	1	0	96	0	27	0	-	-	-
6	5	0	322	0	227	1	-	-	-	-	-
7	1	0	59	0	128	0	-	-	-	-	
S	0.9052		0.8394		0.5848		0.5341		0.4937		0.4884
I_DB	0.4542		0.4166		0.5698		0.5593		0.6607		0.7538

anfitriona, período orbital y otras características físicas. También incluye información sobre la estrella anfitriona, como su nombre, masa y radio. Cabe mencionar que el archivo se actualiza periódicamente a medida que se descubren nuevos exoplanetas. La última actualización del conjunto de datos se hizo en febrero del 2023, el archivo contiene en total 5250 muestras de planetas confirmados pero no etiquetados como similares o no a la Tierra. En la Tabla 1 se describen las 13 características almacenadas de cada exoplaneta conocido hasta el momento.

4.2. Preprocesamiento y limpieza de datos

Este paso es fundamental en el análisis de datos puesto que implica identificar y corregir errores, inconsistencias y datos incompletos en los conjuntos de datos. Por lo tanto, es crucial para garantizar la precisión de los resultados del análisis realizado. A continuación, se explican los pasos involucrados en este proceso sobre los datos del NASA Exoplanet Archive.

Datos nulos. Dentro del conjunto de datos se logró identificar la existencia de datos nulos en ciertas características, la cantidad de datos nulos no era significativa por lo que no se eliminó por completo ninguna característica, en cambio se eliminaron muestras que contenían un dato nulo en alguna de sus características. Una vez realizado este proceso el conjunto de datos pasó de tener 5250 muestras a tener 4765 (el 90.76 % del conjunto original).

Transformación de datos. En el conjunto de datos se identificó que los valores de las características `mass_multiplier` y `radius_multiplier` poseen una escala de medición distinta, específicamente la medición de la multiplicidad de la masa y el radio se hizo con respecto a los planetas Tierra y Júpiter. Por esta razón, todas las muestras cuya multiplicidad de masa y radio fueron medidas con respecto a Júpiter se transformaron para que la escala sea con respecto a la Tierra. Mediante una investigación en [8] y en [10], la masa de Júpiter es 318 veces mayor que la masa de la Tierra, mientras que el radio de Júpiter es 11 veces mayor que el de la Tierra.

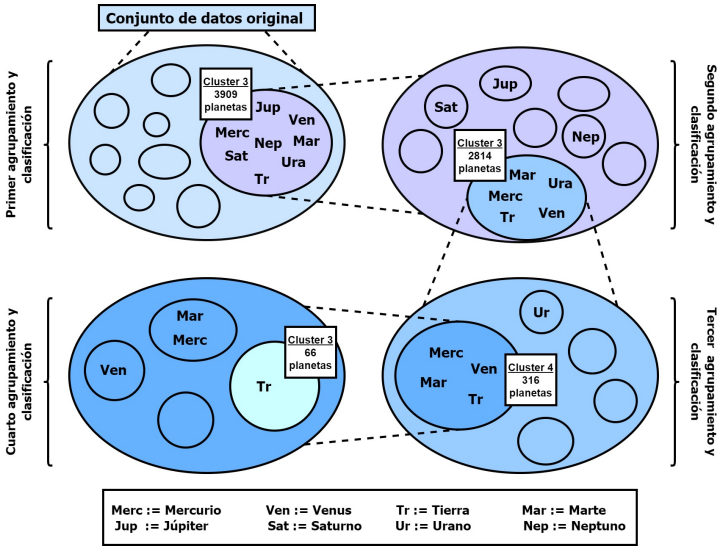

Fig. 1. Resultados obtenidos con agglomerative clustering.

Considerando esta información, todos los valores de `mass_multiplier` y `radius_multiplier` que se midieron con respecto a Júpiter, se multiplicaron por 318 y 11, respectivamente.

Eliminación de variables. En el conjunto de datos existen tanto datos numéricos como categóricos, pero, las técnicas de aprendizaje automático aquí utilizadas no permiten trabajar con variables alfanuméricas, por lo que, se descartan las variables `mass_wrt`, `radius_wrt` y `detection_method`; es importante señalar que dichas variables no representan una característica física de los exoplanetas, lo cual es importante para el propósito de este trabajo. Por otro lado, la característica `discovery_year` pese a ser una variable numérica, tampoco representa una característica física de los explanetas por lo que, también se descartó.

Además se eliminó la variable `planet_type`, ya que esta etiqueta no corresponde a la que se desea predecir. Más bien, es el tipo de exoplaneta derivado de los planetas del sistema solar y categoriza los exoplanetas en los siguientes tipos: gigante gaseoso, neptuniano, supertierra y terrestre. Otra de las tareas comunes en la limpieza de datos es la eliminación de valores atípicos (outliers), sin embargo, se ha optado por omitir esta tarea para tomar en cuenta todos los posibles valores encontrados y analizar e interpretar los resultados que se obtengan en el entrenamiento del modelo.

Selección de variables. La información en [9] influye al momento de seleccionar las variables más apropiadas para la experimentación. Por lo que se considero lo siguiente. La distancia de un planeta a su Sol o estrella anfitriona es un factor fundamental que afecta una amplia gama de características y procesos planetarios, desde el clima y la habitabilidad hasta la dinámica orbital y la exploración espacial; es un aspecto central en la comprensión del cosmos y en la búsqueda de vida en otros mundos. En este sentido, la variable `orbital_radius` resulta ser una característica importante a considerar para el entrenamiento del modelo.

Caso contrario es el de la variable `distance` la cual no proporciona información comparable con respecto a la Tierra ya que esta variable en si misma ya está comparada con respecto a la Tierra. Lo mismo sucede con la variable `stellar_magnitude`, la cual cuantifica el brillo de una estrella o cuerpo celeste observado desde la Tierra, por lo que, convendría no considerar esta variable para el entrenamiento.

Por todo lo anterior, las variables más relevantes para realizar la experimentación son: `mass_multiplier`, `radius_multiplier`, `orbital_radius` y `eccentricity`. La masa y el radio de los exoplanetas se miden en kg. y km., respectivamente, pero los dos primeros datos representan unidades de multiplicidad (números reales no negativos), el radio orbital es medido en AU (unidad astronómica, 1 AU = 149 597 870 700 metros) y la excentricidad es un valor entre 0 y 1.

Conjunto de planetas. Para la parte de aprendizaje supervisado se creó un segundo conjunto de datos recopilando información sobre los planetas del sistema solar, información correspondiente a las cuatro características seleccionadas para el agrupamiento. Los datos de estos planetas muestran la increíble diversidad del sistema solar, adicionalmente, proporcionan una variedad de tipos de planetas que posiblemente se pueda encontrar en otras partes del universo observable, por eso al realizar las tareas de aprendizaje no supervisado se pretende que la agrupación de los datos se haga de tal manera que exista un grupo por cada uno de los planetas del sistema solar. En la Tabla 2 se encuentran las características de los ocho planetas del sistema solar.

4.3. Pruebas

Una vez realizada la limpieza de datos y la selección de características que se utilizarán en los experimentos, se probará la metodología propuesta. El entrenamiento del modelo consta de dos partes, un algoritmo para agrupamiento y otro para clasificación. Para estas dos partes, se utilizó el conjunto de datos de exoplanetas confirmados por la NASA y el conjunto de datos de planetas del sistema solar, respectivamente. Para la parte de aprendizaje no supervisado, se utilizaron los métodos de k-means++, DBSCAN, agglomerative clustering y divisive clustering. Mientras que, para la parte del aprendizaje supervisado, se utilizó k-nn.

Modelo de aprendizaje propuesto. En el algoritmo 1 se muestra el proceso del modelo empleado para encontrar el grupo de planetas más parecidos a la Tierra, siendo X el conjunto de exoplanetas, $C_1 \cup \cdots \cup C_M = \cup\, C \subseteq X$ donde $C = \{C_1, \ldots, C_M\}$ son los grupos de exoplanetas y $P = \{p_1, \ldots, p_M\}$ el grupo de planetas del sistema solar. Inicialmente la cardinalidad M de los conjuntos, C y P, es ocho porque son ocho los planetas del sistema solar.

Durante las pruebas, en el paso 3, se usaron cada uno de los algoritmos de agrupamiento presentados anteriormente. Este paso devuelve los grupos que se forman con los datos de exoplanetas. Una vez creados los clústeres se usó el algoritmo k-nn, paso 5, para determinar a qué grupo pertenece cada uno de los planetas del sistema solar. Este mismo algoritmo de clasificación se utilizó con todas las técnicas de agrupamiento utilizadas. El valor de salida al clasificar los planetas corresponde a la etiqueta del cluster al que pertenecen los planetas del sistema solar de acuerdo con el resultado del agrupamiento.

En el paso 6 se identifica el cluster en el cual se clasificó la Tierra. Dentro del sistema solar hay planetas que no son muy diferentes a la Tierra (en cuanto a las características que se están considerando) por lo que, es de esperar que en la etapa de clasificación más de un planeta quede dentro del mismo grupo que la Tierra, lo cual se determina en el paso 9.

Finalmente, en el paso 10, se verifica si en el cluster donde se clasificó la Tierra éste fue el único planeta (de los 8 del sistema solar) en clasificarse ahí, si es así entonces se termina el proceso y dicho cluster corresponde al de los planetas más similares a la Tierra, paso 11; en caso contrario el proceso se repite sobre el grupo donde se encuentre clasificada la Tierra, para ello, se actualiza el conjunto de exoplanetas y el conjunto de planetas del sistema solar, pasos 13 y 14, respectivamente. En resumen, el proceso termina cuando la Tierra es el único planeta del sistema solar que se clasifica dentro de alguno de los clusters creados automáticamente.

Nótese que el algoritmo va filtrando la información para quedarse únicamente con los datos de los exoplanetas y los planetas del sistema solar que se encuentran en el cluster donde se clasifica la Tierra. Esto permitirá llevar a cabo una reducción de candidatos a planetas similares a la Tierra, enfocándose únicamente en este subconjunto de datos y descartando los datos de los demás clusters y planetas que no se clasificaron junto con la Tierra.

5. Análisis y resultados

Con base al objetivo de este trabajo, se implementó el modelo con los diferentes algoritmos de agrupamiento y clasificación mencionados anteriormente para comparar su rendimiento. Para la implementación se uso el lenguaje de programación Python y sus bibliotecas `scikit-learn` para los algoritmos de clustering y `pandas` para el tratamiento de datos. En esta sección se presentan los resultados de las pruebas y su respectivo análisis.

En la Tabla 3 se pueden apreciar los resultados de cada algoritmo de agrupamiento en la primera iteración. Una vez agrupados los datos y después de clasificar los planetas del sistema solar se observó que el cluster en donde se clasificaba la Tierra era "representativo" de los demás en cuanto a la cantidad de exoplanetas que se agrupaban en éste, ya que prácticamente en todos los casos más del $50\,\%$ de los datos se agrupaban en dicho cluster. En la primera iteración, con k-means++ y el parámetro `k=8` (porque son 8 los planetas del sistema solar), la Tierra se clasificó en el cluster etiquetado con 0, pero no sólo la Tierra, sino que todos los demás planetas del sistema solar también se clasificaron ahí.

Claramente este cluster era representativo y al analizarlo se encontró que en él se encontraban 4526 exoplanetas, correspondientes al 95 % de los datos, mientras que en el cluster 5 se agrupó el 4.7 % de los datos y el resto entre los demás clusters. Con DBSCAN, usando los parámetros por defecto `eps=0.5` y `min_samples=6`, después de crear los clusters y de clasificar los planetas del sistema solar, cinco de estos se clasificaron en el cluster con etiqueta 1, incluida la Tierra. La relevancia de este cluster es que en él se agruparon 3002 exoplanetas, que corresponden al 63 % de los datos disponibles.

Con agglomerative clustering estableciendo el parámetro de `n_clusters = 8`, todos los planetas del sistema solar se clasificaron dentro del cluster etiquetado con 3, dicho cluster contenía 3909 exoplanetas, el 82 % de los datos totales. Para divisive clustering se mantuvo el mismo parámetro respecto al número de clusters y al igual que agglomerative clustering todos los planetas del sistema solar se clasificaron dentro del cluster etiquetado con 2 que también contenía 3909 exoplanetas, el 82 % de los datos. Por lo anterior, se puede concluir que, en la primera iteración todos los métodos de agrupamiento crearon grupos de tamaños muy desiguales, es decir, grupos desbalanceados (uno muy grande y otros demasiado pequeños).

Ahora se analizarán los clusters representativos. Es evidente que k-means++ fue el algoritmo que agrupó más exoplanetas en su primer cluster representativo, por lo que se compararon los datos del cluster representativo de cada algoritmo con respecto al de k-means++. En la comparación se encontró que los 3002 datos que agrupó DBSCAN también los agrupó k-means++ en un mismo grupo. De igual manera, los 3909 datos que agruparon agglomerative clustering y divisive clustering en un mismo grupo, forman parte de los 4526 datos que agrupó k-means++ en su cluster representativo. Por otra parte, agglomerative clustering y divisive clustering son algoritmos jerárquicos, y dado que en sus clusters representativos se agruparon la misma cantidad de datos, estos se compararon y se encontró que eran iguales en cuanto a los datos que agruparon en dicho cluster. Por lo tanto, en la primera iteración, todos los clusters representativos de los distintos métodos eran similares respecto a los datos que contenían.

A partir de los resultados obtenidos con los diferentes algoritmos de agrupamiento se pudo ver que el algoritmo agglomerative clustering junto con divisive clustering tuvieron el mejor desempeño. A continuación, se describe el proceso de entrenamiento del modelo sólo con agglomerative clustering (en la parte de agrupamiento del modelo propuesto). En la figura 1 se puede apreciar como a partir del conjunto de datos original, se van formando subconjuntos en los cuales se clasifican los planetas del sistema solar. Como se puede ver, se necesitó repetir un total de 4 veces el proceso de agrupar y clasificar hasta conseguir que la Tierra quedará en un cluster con 66 elementos del conjunto original sin ningún otro planeta del sistema solar.

Los resultados de cada iteración se muestran en la Tabla 4, donde cada columna se subdivide en dos columnas, la primera muestra la cantidad de exoplanetas agrupados (EA) en el cluster correspondiente, mientras que la segunda muestra el número de planetas del sistema solar clasificados (PC) en dicho cluster. Además, se incluyen los valores de las medidas de calidad del agrupamiento, el coeficiente de silueta S y el índice de Davies Bouldin I_{DB}. Esto permite concluir que, para el método aglomerativo, aunque la calidad del agrupamiento disminuía (gradualmente) los grupos eran balanceados lo que disminuyó el número de iteraciones necesarias para encontrar el grupo con exoplanetas similares exclusivamente a la Tierra.

Cabe mencionar que los mejores resultados con agglomerative clustering se obtuvieron al mantener el parámetro $\mathtt{k} = 10$ fijo en el método de clasificación k-nn. Mediante la experimentación se observó que al variar este parámetro según el tamaño del cluster (inicialmente 100, luego 50, 20 y 10, respecto a la iteración) el número de iteraciones realizadas aumentaba a 5 y el tamaño del grupo de planetas candidatos a ser similares a la Tierra incrementaba a 86.

Los resultados de k-means++ presentados en la Tabla 5 muestran que la calidad de los clusters disminuye con cada iteración y el tamaño de los grupos sigue desbalanceado (sus clusters representativos siguen agrupando una gran cantidad de exoplanetas), lo cuál indica que se requieren de más iteraciones para obtener el resultado deseado. Además, no es posible determinar con exactitud cuantas iteraciones más se requieren para obtener resultados como los del agrupamiento jerárquico. En cuanto a DBSCAN, en la segunda iteración, el algoritmo ya solo generaba un cluster, claramente no proporcionó los resultados esperados.

6. Conclusiones y trabajo futuro

Para la clasificación, el algoritmo k-nn proporcionó buenos resultados. En tanto que, para el agrupamiento los modelos jerárquicos, los métodos agglomerative clustering y divisive clustering, obtuvieron en general los mejores resultados pero, los resultados de k-means++ y DBSCAN no fueron nada esperanzadores. Es evidente que el mejor modelo de aprendizaje se consigue con agglomerative clustering o divisive clustering y k-nn, juntos logran encontrar un grupo de 66 exoplanetas con características similares a la Tierra de entre 5250 planetas.

Los resultados demuestran la importancia de seleccionar técnicas de aprendizaje automático apropiadas para el problema específico que se está abordando, ciertos algoritmos pueden ser más adecuados que otros para detectar patrones en los datos. El desempeño de los algoritmos también puede indicar qué características o variables son más relevantes para identificar exoplanetas parecidos a la Tierra. Para un trabajo futuro, se podría realizar una validación adicional de los resultados utilizando diferentes conjuntos de datos.

Adicionalmente, intentar mejorar la selección de características relevantes para encontrar exoplanetas similares a la Tierra. Finalmente, aunque se ha identificado un algoritmo de agrupamiento efectivo, aún es posible explorar otros algoritmos de agrupamiento y de clasificación para ver si alguno de ellos mejora los resultados. En definitiva, se ha demostrado que la tarea de descubrir planetas fuera del sistema solar y eventualmente encontrar un exoplaneta con condiciones de habitabilidad para los humanos es posible con la ayuda de algoritmos de aprendizaje automático.

Referencias

1. Arthur, D., Vassilvitskii, S.: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics. pp. 1027–1035 (2007)
2. Aggarwal, C. & Reddy, C.: Data clustering: Algorithms and applications. Chapman and Hall/CRC (2013)
3. Herur, A., Tajmohamed, R., Ponsam, J.: Exploring Exoplanets using KNN, Logistic Regression and Decision Trees, 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–7 (2022) doi: 10.1109/ICSES55317.2022.9914278

4. Jin, Y., Yang, L., Chiang, C.: Identifying exoplanets with machine learning methods: A preliminary study. International Journal on Cybernetics & Informatics (IJCI), pp. 31–42 (2022) doi: 10.5121/ijci.2022.110203
5. Martin, E., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), vol. 96, no. 34 pp. 226–231 (1996)
6. Pineda, C.: Aprendizaje automático y profundo en Python. RA-MA (2022)
7. Raschka, S, Mirjalili, V.: Python machine learning: Aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow. Marcombo (2019)
8. Rogers, J.: The giant planet Jupiter, Cambridge University Press, vol. 6 (1995)
9. Perryman, M.: Exoplanet handbook. Cambridge University Press (2018)
10. Rodríguez, H.: Planeta Júpiter: El gigante gaseoso del sistema solar. National Geographic (2023)