Detección de exoplanetas: Interpretando los datos de Kepler

Proyecto final de datos para las asignaturas Preprocesado de datos y Extracción de conocimiento en bases de datos

Juan Manuel Falcón Ramírez

Universidad de La Laguna

Máster universitario en Ciberseguridad e Inteligencia de datos

Curso 2023 - 2024

Índice

1. Introducción

El telescopio espacial Kepler es un satélite construido por la NASA que fue lanzado en 2009. Fue diseñado con el propósito de descubrir planetas similares a la Tierra orbitando la zona habitable de otras estrellas en nuestra región de la Vía Láctea, estudiar su abundancia y determinar las propiedades de las estrellas que tienen sistemas planetarios. El método utilizado para encontrar posibles exoplanetas es la detección por tránsito, pequeñas disminuciones del brillo de una estrella que ocurren cuando un objeto pasa por delante de ella (por ejemplo, un planeta).

En octubre de 2018 fue retirado oficialmente al quedarse sin combustible tras 9 años de recolección de datos, dejando un legado de más de 2600 planetas descubiertos fuera del Sistema solar, siendo algunos de ellos grandes candidatos para albergar vida. El objetivo de este proyecto es hacer uso de los datos recabados por Kepler, para estudiar las características que definen a un exoplaneta y su estrella y crear un modelo utilizando Python que sea capaz de predecir cuando un objeto detectado de ciertas características es un exoplaneta.

2. Datos

Los datos utilizados para la elaboración de esta memoria se han obtenido del repositorio de kaggle correspondiente al enlace https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results, publicados por la NASA. El fichero *cumulative.csv* contiene datos recogidos por Kepler hasta octubre de 2017. Muestra información detallada de los "Kepler objects of interest"(KOI), que son los más de 10000 posibles candidatos que fueron observados utilizando el telescopio. La base de datos contiene 9564 instancias y 50 columnas en total. En la Figura 1 se muestran las primeras instancias del dataframe.

	rowid	kepid	kepoi_name	kepler_name	koi_disposition	koi_pdisposition	koi_score	koi_fpflag_nt	koi_fpflag_ss	koi_fpflag_co	...
0	1	10797460	K00752.01	Kepler-227 b	CONFIRMED	CANDIDATE	1.000	0	0	0	...
1	2	10797460	K00752.02	Kepler-227 c	CONFIRMED	CANDIDATE	0.969	0	0	0	...
2	3	10811496	K00753.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	0	1	0	...
3	4	10848459	K00754.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	0	1	0	...
4	5	10854555	K00755.01	Kepler-664 b	CONFIRMED	CANDIDATE	1.000	0	0	0	...
5	6	10872983	K00756.01	Kepler-228 d	CONFIRMED	CANDIDATE	1.000	0	0	0	...
6	7	10872983	K00756.02	Kepler-228 c	CONFIRMED	CANDIDATE	1.000	0	0	0	...
7	8	10872983	K00756.03	Kepler-228 b	CONFIRMED	CANDIDATE	0.992	0	0	0	...
8	9	6721123	K00114.01	NaN	FALSE POSITIVE	FALSE POSITIVE	0.000	0	1	1	...
9	10	10910878	K00757.01	Kepler-229 c	CONFIRMED	CANDIDATE	1.000	0	0	0	...

Figura 1: Archivo de datos *cumulative.csv*.

2.1. Variables de interés

La información sobre qué significa cada campo recogido en el dataset está explicada detalladamente en el NASA Exoplanet Archive, con enlace https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html. De las 50 columnas que contiene la base, las primeras (sin contar los índices `rowid` y `kepid`) muestran información básica sobre el KOI :

- `kepoi_name`: El nombre del objeto de interés identificado por kepler, que inicialmente debe ser consistente con la hipótesis del tránsito planetario.
- `kepler_name`: Nombre que se le asigna a aquellos objetos que han sido confirmados como planetas.
- `koi_disposition`: Como se muestra el objeto en la literatura (exoplaneta confirmado, candidato, falso positivo o no clasificado).
- `koi_pdisposition`: La clasificación que se le asigna al KOI según el análisis de los datos de Kepler (candidato, falso positivo o no clasificado).
- `koi_score`: Valor entre 0 y 1 que indica la confianza de que la disposición del KOI sea correcta. Cuánto más cerca del 1, mayor es la confianza.

Teniendo en cuenta lo anterior, podemos considerar como variables objetivo a predecir tanto `koi_disposition` como `koi_pdisposition`, ambas se analizarán con mayor detenimiento en la siguiente sección. Los campos `koi_tce_plnt_num`, `koi_tce_delivname` y el resto de variables del grupo TCE (Threshold-Crossing Event) indican cómo se realizó una de las pruebas utilizadas durante la búsqueda de planetas en tránsito.

El resto de variables muestran información relacionada con la observación del objeto y su tránsito, como `koi_period`, que indica el intervalo de tiempo que ocurre entre dos tránsitos consecutivos. Se muestran también características sobre el objeto, como `koi_teq`, una aproximación de su temperatura, y algunos parámetros estelares, como `koi_smass` o `koi_sage`, que recogen la masa y edad de la estrella que orbitan. Las columnas acabadas en `_err` indican las incertidumbres positivas y negativas asociadas a su correspondiente variable.

2.2. Preprocesado de los datos

2.2.1. Análisis exploratorio

A continuación, se realizará un breve análisis preliminar para determinar que variables serán de utilidad a la hora de crear el modelo y estudiar los datos. En primer lugar, se estudia la estructura de la base para evitar posibles errores. Todas las variables son numéricas exceptuando 5 variables categóricas: `kepoi_name`, `kepler_name`, `kepoi_disposition`, `kepoi_pdisposition` y `koi_tce_delivname`, por lo que en principio no existen valores de un tipo que no le corresponde a algún campo en particular. Las variables `koi_teq_err1` y `koi_teq_err2` están completamente vacías, así que las eliminaremos. El siguiente paso será decidir cuál será la variable objetivo del modelo.

En la Figura 2 se muestran las variables `koi_disposition` y `koi_pdisposition` clasificadas. En el caso de `koi_disposition`, existen tres clases diferentes, de las cuales 'CONFIRMED' Y 'FALSE POSITIVE' muestran objetos que ya han sido evaluados y verificados, mientras que 'CANDIDATE' hace referencia a los KOI que, en el momento en el que se publicaron los datos, seguían siendo estudiados para determinar su categoría. Por otro lado, el caso de `koi_pdisposition` es más simple, separando únicamente entre candidatos y falsos positivos. Dado que el objetivo propuesto es tratar de deducir que objetos KOI son exoplanetas, nos interesan aquellos que ya han sido confirmados. Por lo tanto, nos interesan las categorías CONFIRMED y FALSE POSITIVE de la variable `koi_disposition`.

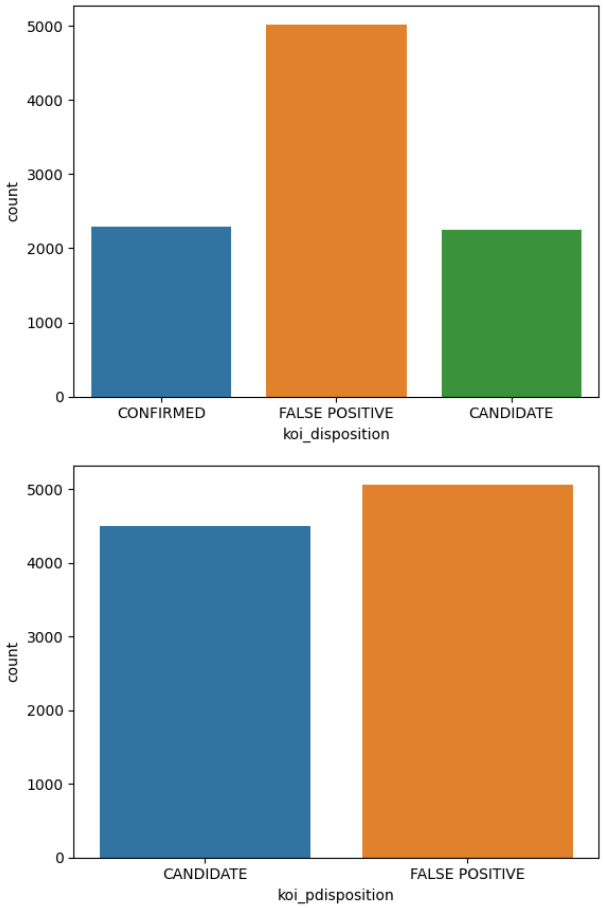

Figura 2: Distribución de clases para `koi_disposition` (arriba) y `koi_pdisposition` (abajo).

Por facilitar el procedimiento, eliminaremos las variables que no serán de utilidad para el proyecto (Figura 3), incluyendo todas aquellas que no aportan información sobre las características del objeto, su tránsito, la estrella que orbitan o su lugar en el cielo. En concreto, se eliminan las variables pertenecidentes a las categorías 'Threshold-Crossing Event', 'Project Disposition Columns' y las columnas identificadoras explicadas en el NASA Exoplanet Archive. El resultado es un dataframe de 35 columnas.

```
Variables eliminadas:  ['rowid', 'kepid', 'kepoi_name', 'kepler_name', 'koi_fpflag_nt']
                       ['koi_fpflag_ss', 'koi_fpflag_ec', 'koi_fpflag_co', 'koi_score', 'koi_pdisposition']
                       ['koi_tce_plnt_num', 'koi_tce_delivname', 'koi_teq_err1', 'koi_teq_err2', 'koi_model_snr']
```

Figura 3: Columnas eliminadas.

2.2.2. Tratamiento de valores nulos

Seguimos el análisis repasando los valores nulos presentes en los datos. En general, la gran mayoría de variables poseen entre 300 y 500 valores nulos, pero como el número total de individuos es 9564, no suponen un problema. Tras comprobar que en todas las variables los valores se encuentran relativamente distribuidos y no se da el caso de que la mayoría de estos estén en los extremos, se utilizará la media para completar los valores faltantes.

2.2.3. Detección de outliers

Debido al tamaño del universo, cada objeto que descubramos en el espacio puede presentar características únicas. Tan sólo con observar el Sistema Solar nos podemos percatar de las grandes diferencias que existen entre los planetas que lo forman. Debido a esto, utilizando métodos de detección de outliers encontraremos más individuos extraños de los que cabría esperar, por lo que en esta sección nos centraremos únicamente en identificar los casos que parezcan erróneos −no los raros− y los eliminaremos.

A la hora de detectar anomalías, resulta de mucha utilidad visualizar los datos en forma de gráficos de cajas, puesto que de esa manera saltan a la vista valores que no parecen ser correctos. Por ejemplo, en la Figura 4 se muestra el caso de la variable `koi_period`, que hace referencia al periodo orbital −el intervalo de tiempo entre dos tránsitos del objeto, lo que tarda en dar la vuelta a su estrella− en unidades de días, y `koi_duration`, la duración del tránsito. Se aprecia claramente que un individuo tiene un periodo orbital considerablemente mayor al resto, concretamente de ∼129996 días (∼356 años terrestres).

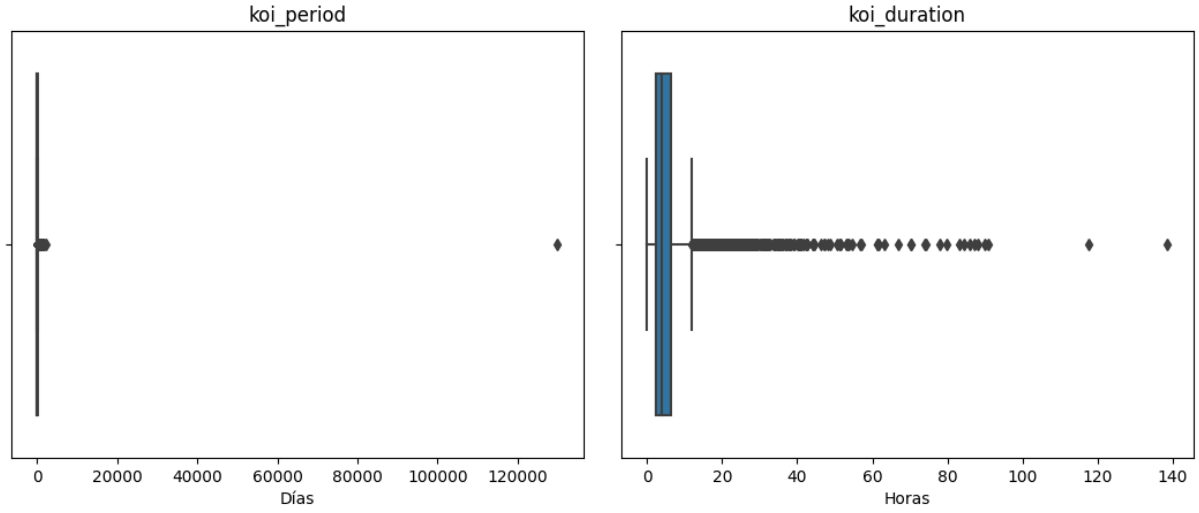

Figura 4: Gráficos de cajas de las variables `koi_period` (izquierda) y `koi_duration` (derecha).

Si observamos la variable `koi_duration`, notaremos que la duración de su tránsito es de 9 horas, lo cual ni tiene mucho sentido ni existe otro individuo que tenga una duración de su órbita compatible con el periodo orbital descrito. La única opción es que el KOI se mueva extremadamente rápido o que se trate de un error. Como referencia, el planeta con el año más largo detectado por tránsito es "EPIC 248847494 b", que tarda 10 años en dar una vuelta alrededor de su estrella. El periodo más largo detectado utilizando otro método sería el de "COCONUTS-2 b", con un año de 1101369.9 años terrestres de duración, pero se trata de un caso muy concreto. Observando que varios de los demás campos también poseen un único individuo que parece estar fuera de lugar, en la Figura 5 se muestra lo que ocurre al eliminarlo.

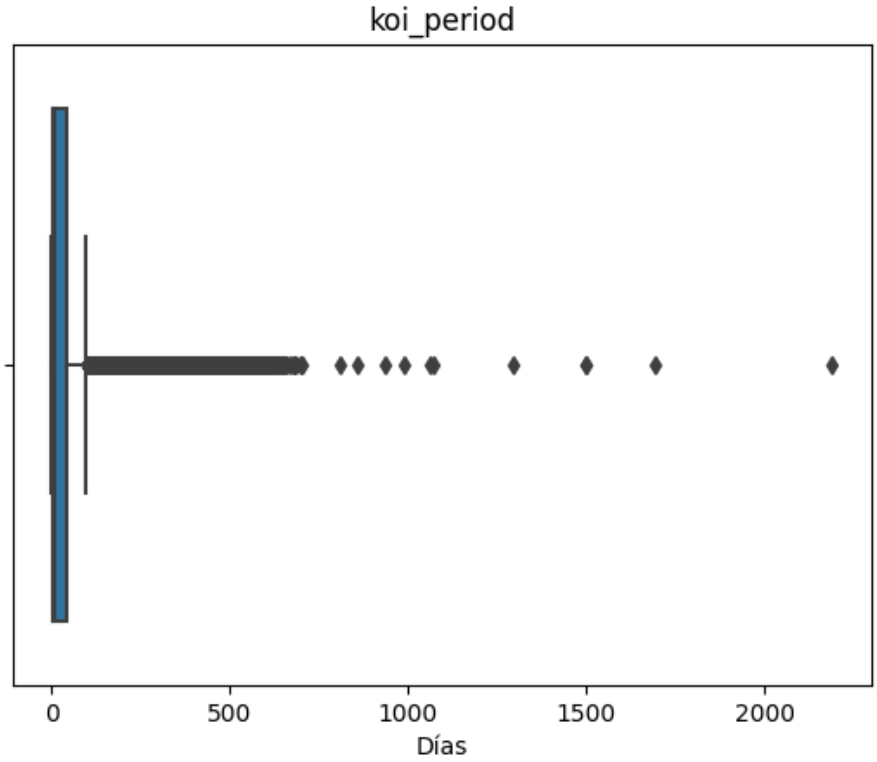

Figura 5: Gráfico de cajas de la variable `koi_period` tras la eliminación del outlier.

El gráfico de cajas de `koi_period` tiene más sentido ahora, presentando valores más alejados de la media pero dentro de lo razonable. Del mismo modo, algunas de las variables dejan de presentar anomalías, pero existen otras que siguen presentando outliers claros, como `koi_depth`. Observando que se trata de un único valor, podemos considerarlo nuevamente como un error en los datos. Por último, siguiendo el mismo procedimiento se eliminan dos individuos con valores anómalos en la columna `koi_insol`.

2.2.4. Estudio de la correlación

Terminamos el preprocesado de los datos realizando un estudio sobre la correlación de las variables que se van a utilizar. En el mapa de calor siguiendo el método de Pearson de la Figura 6, se muestra la correlación de las variables del dataset, excluyendo las variables que representan incertidumbres (las acompañadas por `_err`).

Entre las positivas, destacan `koi_impact`, la proyección en el cielo de la distancia entre el centro de la estrella y el centro del objeto, y `koi_prad`, el radio del KOI. También están relacionadas `koi_time0bk`, la fecha en días julianos baricéntricos en la que se detectó el tránsito, con el periodo orbital. El primer caso lo obviaremos a pesar de presentar una mayor relación, puesto que las dos variables transmiten información importante incluso siendo algo redundante; pero en el segundo caso eliminaremos `koi_time0bk`, dado que la hora de la detección no debería influir al determinar si un objeto es o no un exoplaneta.

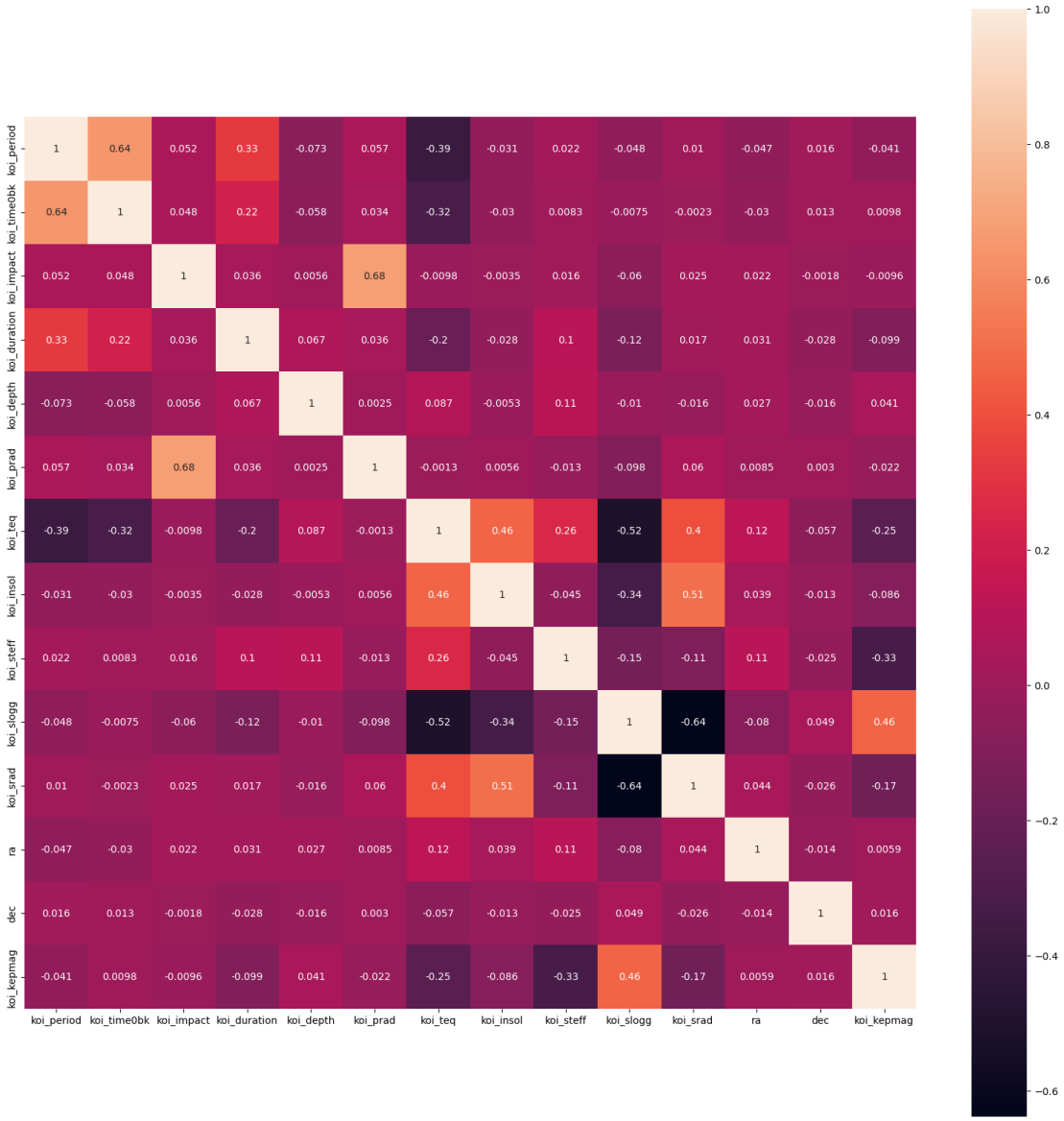

Figura 6: Mapa de correlación.

Por otro lado, encontramos una relación negativa entre las variables `koi_slogg` y `koi_srad`, que describen la gravedad superficial y el radio de la estrella asociada al KOI respectivamente (Figura 7). Esta relación es conocida en astrofísica y puede entenderse fácilmente al comparar estrellas enanas con estrellas gigantes. Las enanas blancas son formaciones estelares extremadamente densas, con un pequeño tamaño dado que se han comprimido debido a su intensa fuerza gravitatoria, mientras que las gigantes rojas son el caso contrario, estrellas menos densas con gravedad menos intensa y, por ende, suelen acabar con un tamaño mayor.

Aprovechando este análisis, podemos ampliar los objetivos del proyecto realizando un agrupamiento no supervisado para determinar el tipo de estrellas con las que estamos tratando, basándonos en los parámetros estelares recogidos en la base de datos. Este proceso se explicará con mayor detalle en la sección dedicada al agrupamiento. En este punto acaba el preprocesado de los datos, con un dataframe resultante de 9560 filas y 34 columnas.

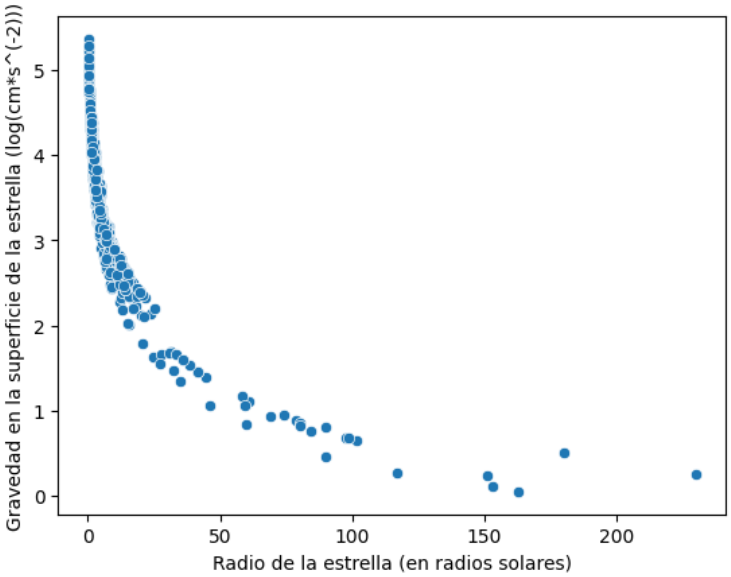

Figura 7: Gravedad en la superficie de la estrella frente a su radio.

3. Clasificación

La finalidad de este proyecto es crear un modelo capaz de predecir qué objetos detectados son exoplanetas. Entre los tres valores de la variable objetivo `koi_disposition`, descartaremos los elementos clasificados como CANDIDATES, puesto que no habían sido verificados como exoplanetas o falsos positivos en el momento de la publicación de los datos. Por lo tanto, el modelo deberá diferenciar aquellos individuos clasificados como CONFIRMED de los clasificados como FALSE POSITIVE conociendo únicamente algunas características físicas, de su estrella, de la observación de su tránsito y de su posición.

Al eliminar a los candidatos, el dataframe resultante cuenta con 7314 individuos de los cuales 5021 son falsos positivos y 2293 son exoplanetas confirmados, existe un claro desbalanceo de clases. El siguiente paso será obtener el conjunto de entrenamiento y de validación a través de la función *train_test_split* del módulo *sklearn*, habiéndose especificado que se utilice un 75 % de los datos para el entrenamiento del modelo.

3.1. Corrección del desbalanceo de clases

El conjunto de entrenamiento resultante está formado por 5485 individuos, de los cuales 3759 están clasificados como falsos positivos y 1726 como exoplanetas. Esta diferencia puede provocar un sesgo en el modelo final e impedir que clasifique correctamente. A modo de ejemplo, si la proporción de clases desbalanceadas —en el conjunto de entrenamiento— fuese 90/10, el modelo afirmará que todos los datos del conjunto de validación son de la clase correspondiente al 90 % y sin clasificar nada obtendrá una precisión muy alta, por lo que pierde su cometido. Lo interesante es obtener un modelo capaz de discernir correctamente ese 10 % de individuos, y para lograrlo es necesario balancear las clases del conjunto de entrenamiento.

La técnica empleada para balancear será el *oversampling* o sobremuestreo. Consiste en repetir aleatoriamente instancias en la clase menos representada hasta lograr un tamaño del subconjunto igual al subconjunto de la clase más representada. En este caso, se realiza

un sobremuestreo del subconjunto de entrenamiento de los exoplanetas confirmados hasta que su tamaño sea de 3759 individuos, obteniendo un conjunto de entrenamiento total de 7518 instancias.

3.2. Normalización de los datos

La normalización de los datos consiste en transformar las variables para que sean más fieles a una distribución normal, lo que se consigue −en este caso− restando la media y dividiendo por la desviación típica, y suele emplearse con el fin de obtener mejores resultados puesto que los datos se moverán en rangos similares. En el programa de Python se ha resuelto esta situación aplicando directamente la función $StandardScaler()$ de *sklearn* a los datos del conjunto de entrenamiento.

3.3. Creación del modelo

Un algoritmo o modelo de clasificación es una técnica de aprendizaje supervisado que utiliza los datos de un conjunto de entrenamiento proporcionado para aprender a diferenciar entre distintas clases. De todos los métodos de clasificación que existen, haremos uso de la regresión logística, el árbol de clasificación, el método Random Forest y el método Näive Bayes, con el fin de valorar los resultados obtenidos por cada uno de ellos.

A la hora de compararlos, analizaremos la información que se obtiene a partir de la matriz de confusión de cada modelo, matriz que muestra los valores acertados y los falsos positivos y negativos. Entre las medidas estadísticas obtenidas, nos centraremos en las siguientes:

- La exactitud ("accuracy" en inglés) se refiere a lo cerca que está el resultado de una medición del valor verdadero. Se representa como la proporción de resultados verdaderos (verdaderos positivos y verdaderos negativos) dividido entre el número total de casos examinados.
- La precisión ("precision" en inglés) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. La precisión positiva se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos como falsos positivos).
- La sensitividad o 'recall' es la proporción de casos positivos que fueron correctamente identificados por el algoritmo.
- La especificidad indica los casos negativos que el algoritmo ha clasificado correctamente.

Existen más métricas interesantes, como la puntuación F1, que tiene en cuenta tanto la precisión como el recall. No obstante, no será necesaria para estudiar los cuatro modelos obtenidos. A continuación, las Tablas 1, 2, 3 y 4 presentan la matriz de confusión y los parámetros estadísticos obtenidos para cada modelo, denominando al grupo CONFIRMED (la clase positiva) con la letra C y al FALSE POSITIVE (la clase negativa) con la letra F.

3.3.1. Regresión logística

La regresión logística es utilizada para problemas de clasificación binaria, donde el objetivo es predecir si una instancia pertenece a una de dos clases posibles, como en este caso. Utiliza la función logística para modelar la relación entre las variables de entrada y se entrena ajustando los coeficientes de la función para que se asemeje a los datos observados. Después del entrenamiento, el modelo asigna una probabilidad a cada instancia y se elige un umbral, considerándose que una instancia pertenece a la clase positiva si la probabilidad predicha es mayor que el umbral, y a la clase negativa si es menor.

	C	F	TOTAL
$\hat{C}$	535	32	567
$\hat{F}$	148	1114	1262
TOTAL	683	1146	1829

Medidas	
Accuracy	0.90
Positive precision	0.78
Negative precision	0.97
Recall	0.94
Specificity	0.88

Cuadro 1: Resultados de la regresión logística.

Observando la matriz de confusión, la exactitud o precisión total es aceptable. Sin embargo, interesa considerar sobre todo los parámetros relacionados con la clase positiva −nuestro objetivo final sigue siendo identificar exoplanetas, no falsos exoplanetas− y comprobar la eficiencia del modelo en ese aspecto. El modelo tiene una precisión baja y un recall alto, lo que quiere decir que el algoritmo detecta bien la clase, pero incluye también individuos de la otra clase.

3.3.2. Árbol de clasificación

Durante el entrenamiento de un árbol de clasificación, se divide el conjunto de entrenamiento en subconjuntos más pequeños basándose en características, construyendo un árbol que representa estas divisiones. La selección de características y umbrales se realiza para maximizar la separación entre clases. Después del entrenamiento, se clasifican las instancias siguiendo las ramas del árbol.

	C	F	TOTAL
$\hat{C}$	484	83	567
$\hat{F}$	84	1178	1262
TOTAL	568	1261	1829

Medidas	
Accuracy	0.91
Positive precision	0.85
Negative precision	0.93
Recall	0.85
Specificity	0.93

Cuadro 2: Resultados del árbol de clasificación.

El árbol de clasificación muestra mejores resultados en general en lo que a clasificar se refiere. A pesar de que unos 80 elementos de cada clase están intercambiados, el porcentaje de acierto es relativamente alto. La precisión positiva es mayor, dado que los positivos predichos contienen menos cantidad de negativos; mientras que el recall es menor, dado que no se detectan tantos positivos reales como en el caso anterior.

3.3.3. Random Forest

El método Random Forest se basa en la construcción de múltiples árboles de decisión durante el entrenamiento. Cada árbol se entrena en un subconjunto aleatorio del conjunto de datos y utiliza características seleccionadas al azar en cada división. Luego, durante la predicción, los resultados de los múltiples árboles se combinan para obtener una predicción más robusta y generalizable. Es especialmente efectivo para evitar el sobreajuste.

	C	F	TOTAL
$\hat{C}$	514	53	567
$\hat{F}$	64	1198	1262
TOTAL	578	1251	1829

Medidas	
Accuracy	0.94
Positive precision	0.89
Negative precision	0.96
Recall	0.91
Specificity	0.95

Cuadro 3: Resultados del clasificador Random Forest.

De entre todos los métodos estudiados, el Random Forest sin duda presenta los mejores resultados. Se podría plantear como una mejora directa del árbol de clasificación, con mayor accuracy, precisión positiva y recall. En general, el modelo predice mejor los elementos de la clase positiva y se reducen los falsos positivos y negativos.

3.3.4. Näive Bayes

El modelo Näive Bayes basa sus predicciones en el teorema de Bayes, asumiendo la independencia condicional entre las características. Durante el entrenamiento, el modelo calcula las probabilidades condicionales de cada característica dada la clase. Luego, durante la predicción, utiliza estas probabilidades para determinar la clase más probable para una nueva instancia.

	C	F	TOTAL
$\hat{C}$	559	8	567
$\hat{F}$	428	834	1262
TOTAL	987	842	1829

Medidas	
Accuracy	0.76
Positive precision	0.57
Negative precision	0.99
Recall	0.99
Specificity	0.66

Cuadro 4: Resultados del clasificador Näive Bayes.

Debido a que las variables no son del todo independientes (ver la Figura 6), el modelo Näive Bayes no proporciona buenos resultados. A pesar de que el recall es muy alto y que la precisión negativa es muy buena, ya se comentó anteriormente que lo que buscamos en el modelo es que sea capaz de diferenciar los exoplanetas reales. En este caso, más de la mitad del conjunto de validación de ha considerado como clase positiva, cuando en realidad solo 567 individuos son positivos.

3.4. Clasificación de los candidatos

Para concluir la clasificación y quedando claro que el modelo más eficiente ha sido el Random Forest, podemos ponerlo a prueba con los individuos de la clase CANDIDATES

que quedaron eliminados al principio de la sección. Estos objetos siguen en fase de estudio, por lo que no se ha determinado si son realmente exoplanetas o no. Lógicamente, no podemos confirmar que el modelo haya o no acertado con la predicción de estos planetas dado que no tenemos con qué contrastarlo, pero puede servir a modo de estimación.

Hay un total de 2246 individuos clasificados como CANDIDATES. Al aplicar el modelo Random Forest, se han confirmado como exoplanetas a 825 candidatos y como falsos positivos a 1421 candidatos. El porcentaje de exoplanetas confirmados predichos es del 36.7 %, mientras que el porcentaje de exoplanetas confirmados en el conjunto original (eliminando los candidatos de la ecuación) es del 31.4 %. En teoría, no podemos garantizar la eficiencia del modelo basándonos en esta predicción, dado que se puede dar el caso de que todos los planetas candidatos sean confirmados o rechazados como exoplanetas, por ejemplo. Sin embargo y a pesar de dicha posibilidad, que el modelo mantenga un porcentaje similar al original es un buen indicio de su capacidad para predecir exoplanetas.

4. Agrupamiento

El agrupamiento o 'clustering' es una técnica de aprendizaje no supervisado de mucha utilidad cuando se tiene un conjunto de datos y se desea encontrar patrones o estructuras ocultas sin etiquetas predefinidas. Es un buen método para reducir la dimensionalidad del problema y segmentar los datos, dado que estos quedan representados de forma más compacta. En el caso de nuestros datos, el enfoque del agrupamiento serán los parámetros estelares `koi_slogg`, `koi_steff` y `koi_srad` (gravedad, temperatura efectiva y radio de la estrella respectivamente), con el fin de encontrar alguna relación entre los exoplanetas confirmados y las estrellas que orbitan.

4.1. K-Means

Se hará uso del algoritmo K-means, que consiste en asignar cada punto de datos a un clúster de manera que la suma de las distancias al cuadrado entre los puntos y el centroide (punto medio) del clúster sea mínima. En la Figura 8 se observa que, aplicando el método del codo, el número óptimo de clústers es 4.

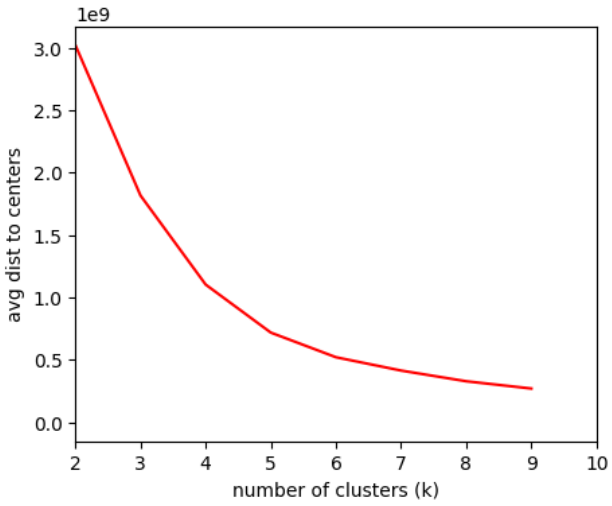

Figura 8: Método del codo.

El resultado de aplicar el algoritmo es la creación de 4 clústers. En la Figura 9 se muestran el total de individuos por clúster: el 0 tiene 270 instancias asociadas, el 1 tiene 4564, el 2 tiene 978 y el 3 tiene 3798.

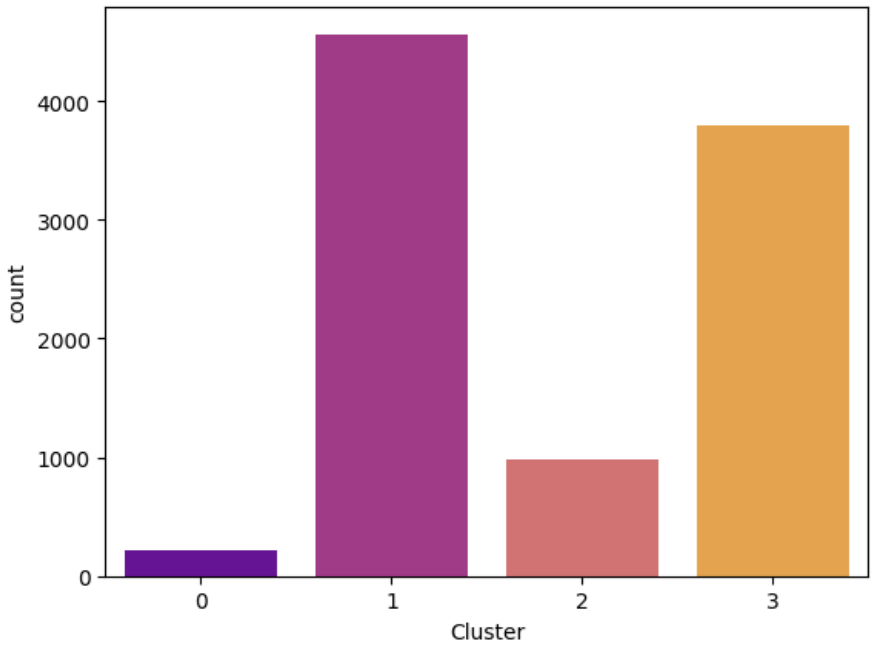

Figura 9: Gráfico de barras que cuenta el número de individuos asociado a cada clúster.

Las Figuras 10 y 11 muestran dos gráficos de dispersión de la variable `koi_steff` donde los grupos han quedado bien diferenciados. El primero de ellos utiliza la magnitud de Kepler del objeto KOI, siendo la magnitud una medida adimensional que indica qué tan brillante es un objeto visto desde La Tierra.

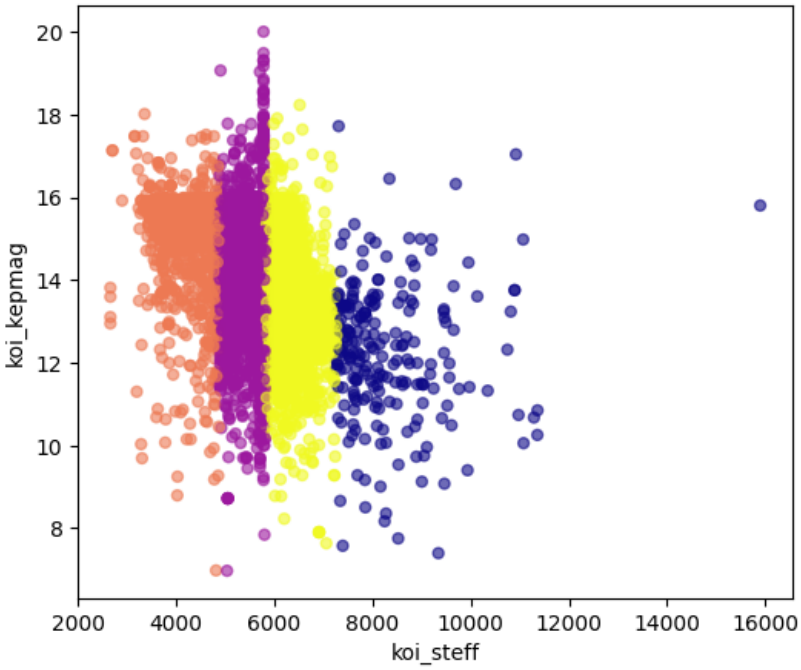

Figura 10: Magnitud de Kepler frente a temperatura efectiva de la estrella (Kelvin).

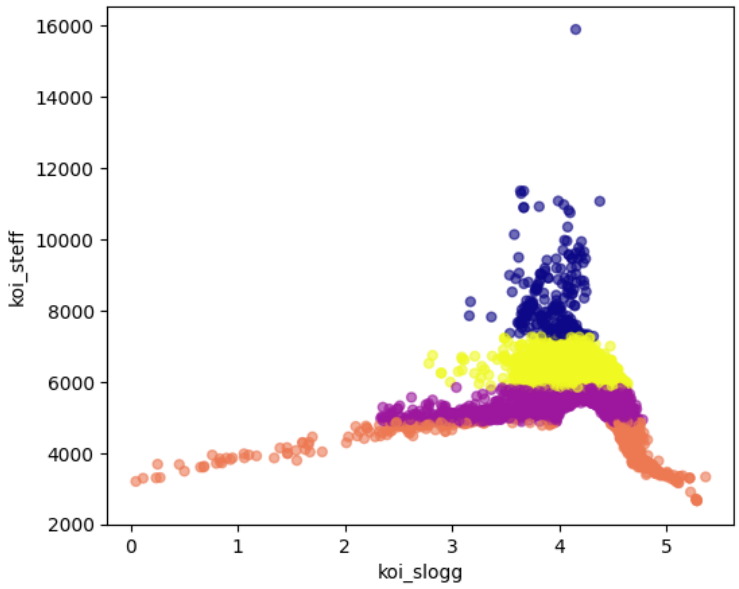

Figura 11: Temperatura efectiva de la estrella (Kelvin) frente a la gravedad en la superficie de la estrella.

4.2. Agrupación por tipo espectral de las estrellas

Representaciones gráficas utilizando otras variables, como la de la Figura 12 que imita a la Figura 7, no muestran una separación apreciable de los clústers. La temperatura efectiva parece ser el patrón principal que ha seguido el modelo para encontrar grupos.

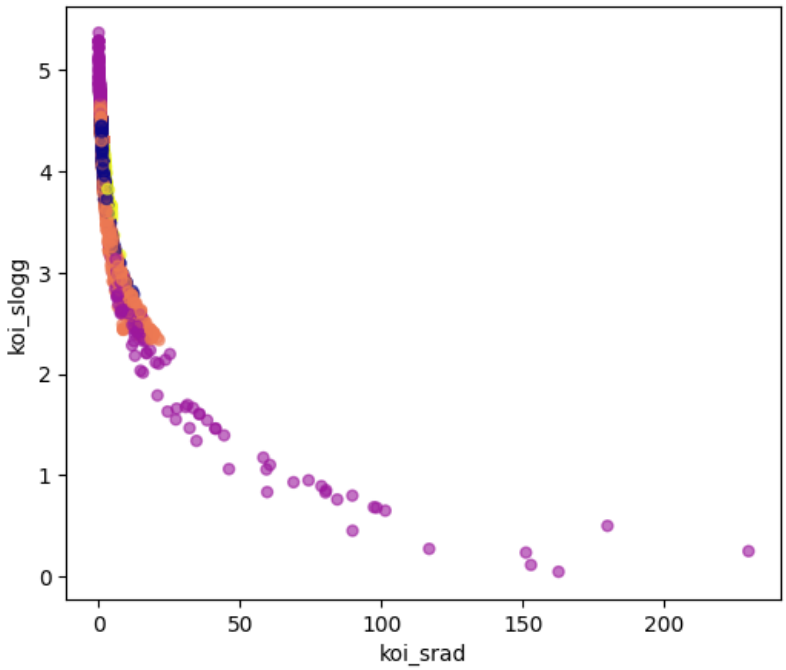

Figura 12: Gravedad en la superficie de la estrella frente a su radio, coloreado en función del clúster al que pertenece el individuo.

Teniendo en cuenta lo anterior, podemos organizar los datos de manera que representen los tipos de estrellas conocidos según su espectro. Esta clasificación se obtiene de analizar el espectro de la estrella y comprobar en qué longitud de onda se encuentra su pico de radiación máxima, pero como no disponemos de dicha información, utilizaremos la temperatura efectiva en su lugar, que está directamente relacionada. De forma resumida, cuanto más caliente sea una estrella, emite radiación electromagnética de mayor energía y más azul será su color (menor longitud de onda). Por esta razón las estrellas más brillantes son azules y las menos brillantes son rojas. Se clasifican por letras de la siguiente manera:

- Tipo O: $>$ 30000 Kelvin, azul violeta.
- Tipo B: 10000-30000 Kelvin, azul blanco.
- Tipo A: 7500-10000 Kelvin, blanco.
- Tipo F: 6000-7500 Kelvin, amarillo blanco.
- Tipo G: 5000-6000 Kelvin, amarillo (tipo del Sol).
- Tipo K: 3500-5000 Kelvin, naranja.
- Tipo M: $<$ 3500 Kelvin, rojo anaranjado.

A continuación, creamos una nueva clase del dataframe con el tipo espectral de la estrella. El objetivo será comprobar si el agrupamiento realizado anteriormente sigue este patrón espectral. En la Figura 13 se aprecia como la mayoría de objetos detectados se encuentran orbitando estrellas tipo G, F y K, de las cuales G tiene la mayor parte de los exoplanetas confirmados. Esto tiene sentido, dado que uno de los propósitos de buscar exoplanetas es encontrar vida, y qué mejor sitio hay para empezar a buscar que planetas con estrellas similares a nuestro Sol.

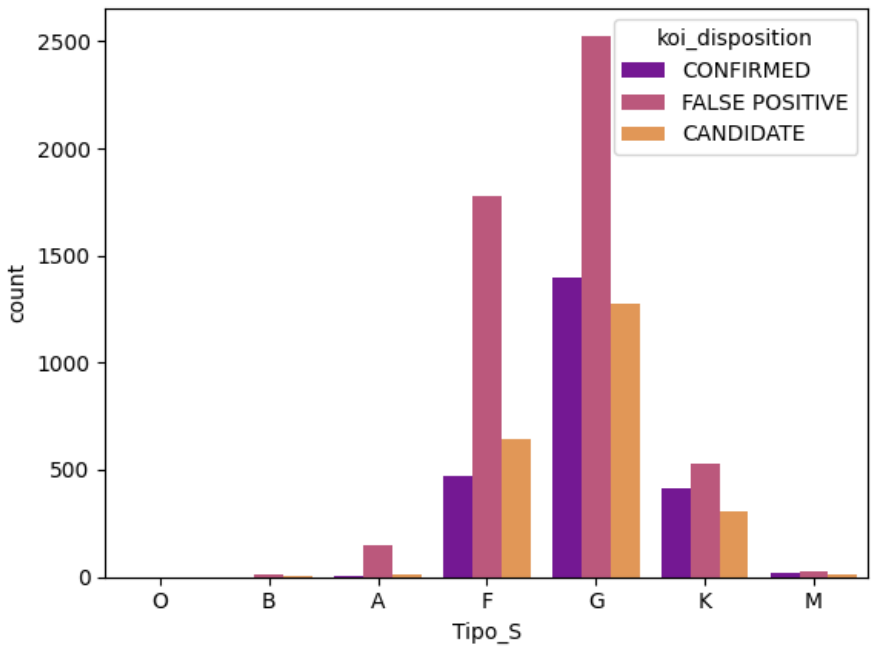

Figura 13: Objetos KOI agrupados en función del tipo espectral de su estrella.

Si representamos los tipos espectrales en función del clúster al que pertenecen −y el patrón previsto se cumple− deberíamos encontrar un gráfico que comparta algunas similitudes con el anterior: un clúster que contenga las estrellas tipo G, otro con la F, otro con la K y un último grupo con los tipos restantes. En la Figura 14, se comprueba que el resultado no dista de lo esperado.

El clúster añil contiene la mayoría de estrellas de la categoría F y una quinta parte de las estrellas tipo G. El clúster azul claro contiene casi todas las estrellas G y una pequeña parte de las K. El clúster rosado contiene el resto de las estrellas K y las pocas que caen en la categoría M. Finalmente, el clúster rojo abarca los tipos de estrella restantes A y B, y una pequeña cantidad de estrellas F.

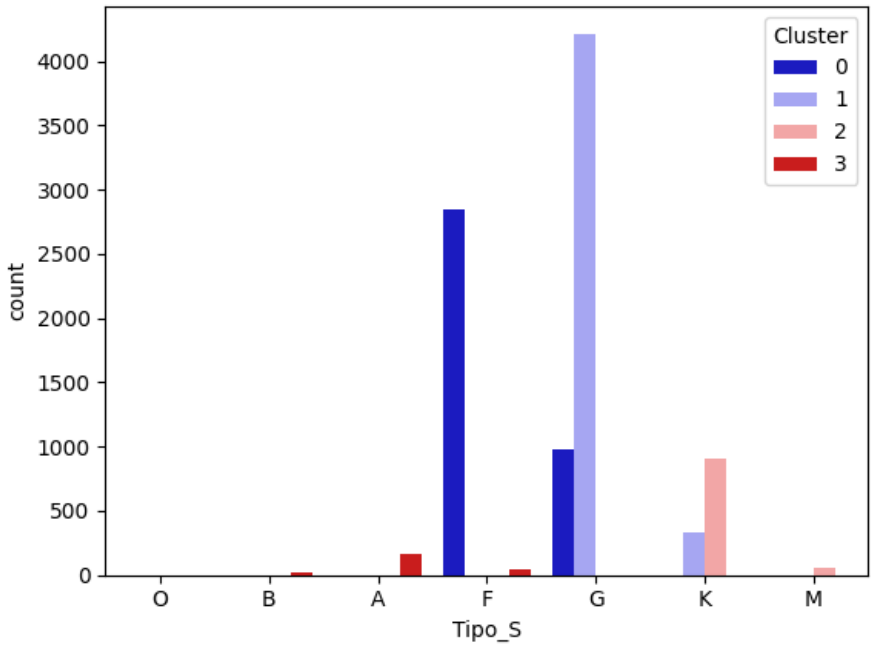

Figura 14: Objetos KOI agrupados en función del tipo espectral de su estrella.

El agrupamiento no ha sido exacto, pero el patrón existe. En el caso de que no conociéramos la clasificación por tipo espectral de las estrellas, este análisis nos hubiese permitido ver que por los menos existen 4 tipos de estrellas diferentes en función de su temperatura, y que tienen cierto peso en los datos. Nos hubiese dado una pequeña pista de por dónde empezar a buscar.

5. Conclusiones

Una de las cuestiones científicas sin resolver más importantes es si existe o no vida fuera de La Tierra. El estudio y la detección de exoplanetas permite investigar individuos con características similares a nuestro planeta, aquellos con mayor probabilidad de albergar vida en las condiciones que conocemos. Este proyecto no es más que una pequeña demostración de lo que la ciencia de datos puede aportar a la búsqueda de nuevos mundos: las técnicas de aprendizaje no supervisado permiten observar patrones escondidos entre los datos que nos lleven a elegir mejor los candidatos a estudiar, y por otro lado las técnicas de clasificación nos permiten predecir cuáles de esos candidatos son exoplanetas en

realidad y cuáles deben ser el foco de investigación. Todo ello combinado permite reducir considerablemente el tiempo transcurrido entre descubrimientos, algo de especial relevancia en un campo como la astrofísica, que estudia los casi infinitos objetos que viajan por el universo.

Los resultados obtenidos son satisfactorios. A pesar de haber eliminado una parte de las columnas originales, el modelo de clasificación es capaz de diferenciar exoplanetas con un alto porcentaje de acierto. En una observación a pequeña escala seguirá siendo mejor el ojo de un experto, pero si aumentase el número de candidatos de forma drástica, aplicar el modelo ahorraría una gran cantidad de trabajo, tiempo y dinero. En posibles líneas futuras de este trabajo, se podría mejorar el algoritmo con nuevas variables que aporten otro tipo de información que pueda ser relevante. También podría enfocarse la detección de exoplanetas en la observación de estrellas del tipo F, G y K, ya que contienen la mayor cantidad de exoplanetas confirmados y candidatos. Podrían estudiarse también otros métodos de agrupamiento, con el fin de encontrar patrones en las características del propio exoplaneta que ayuden a reducir la cantidad de variables necesarias para el entrenamiento del modelo sin reducir su eficiencia.

Otra forma de avanzar en la investigación podría ser ampliar el área de la búsqueda. La siguiente Figura 15 muestra la pequeña zona del cielo donde se realizó la observación, con los clústers identificados anteriormente relativamente bien distribuidos. Es una realidad que existen áreas del universo con mayor concentración de elementos, y en algunos casos la distribución de estos no es equitativa. Tal vez la clave para encontrar nuevos exoplanetas radique en encontrar un apartado del cielo con mayor densidad de estrellas tipo G, pudiéndose obtener esta información de un estudio de galaxias. Las galaxias de un color más rojizo indican poca presencia de estrellas calientes, por lo que aumentaría el porcentaje de estrellas similares al Sol.

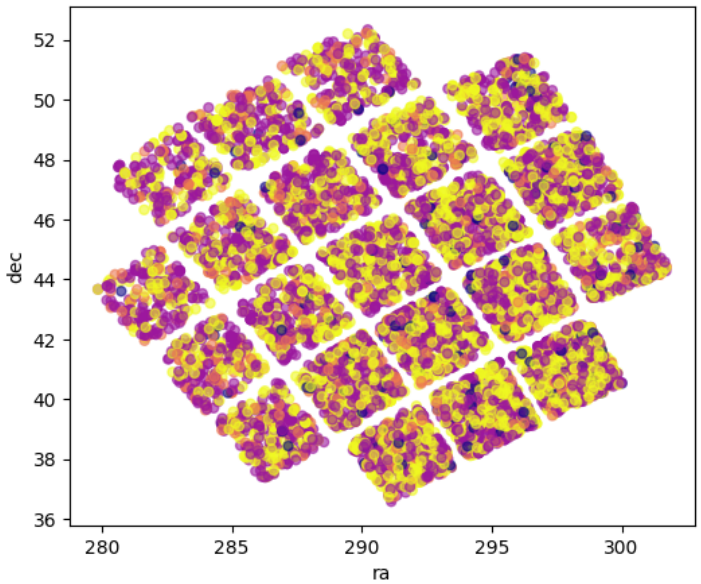

Figura 15: Declinación frente a ascención recta (en grados sexagesimales).

6. Material consultado

Base de datos, publicada por la NASA. https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results

Misión espacial Kepler, descripción de la misión, NASA. https://science.nasa.gov/mission/kepler/in-depth

Información sobre exoplanetas, NASA. https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/overview/

NASA Exoplanet Archive. Información sobre las variables columna de la base de datos de Kepler. https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

Tipos de estrellas, Las Cumbres Observatory. https://lco.global/spacebook/stars/types-stars/

Notebook del usuario Iftesha Najnin "ifteshanajnin", publicado en kaggle. Se estudia el mismo conjunto de datos, enfocándose en el preprocesado y la clasificación. https://www.kaggle.com/code/ifteshanajnin/exoplanet-detection-on-kepler-data

Notebook del usuario Gabriel Atkin "gcdatkin", publicado en kaggle. Breve cuaderno con métodos de clasificación de los datos. https://www.kaggle.com/code/gcdatkin/exoplanet-identification-classification

Información sobre EPIC 248847494 b, NASA. https://exoplanets.nasa.gov/exoplanet-catalog/7435/epic-248847494-b/

Información sobre COCONUTS-2 b, NASA. https://exoplanets.nasa.gov/exoplanet-catalog/7945/coconuts-2-b/