

PROYECTO FINAL

REBECA AMOR, PALOMA MAREQUE, LEANDRO SCANIELLO

Clasificación estelar: Clasificación de estrellas, galaxias y cuásares



INTRODUCCIÓN

Este dataset tiene como objetivo clasificar estrellas, galaxias y cuásares en función de sus características espectrales.

Descripción

El proyecto tiene como objetivo desarrollar un modelo de aprendizaje automático capaz de clasificar estrellas, galaxias y cuásares. Utilizando un conjunto de datos que contiene información astronómica, se entrenará el modelo para distinguir entre los diferentes tipos de objetos celestes. Mediante el uso de técnicas de aprendizaje automático, se busca lograr una clasificación precisa y eficiente que pueda ser aplicada a grandes conjuntos de datos astronómicos.

Utilidades principales

Este modelo entrenado para clasificar estrellas, galaxias y cuásares puede ser útil en varios aspectos como:

Investigación astronómica

La clasificación precisa de objetos astronómicos es fundamental para comprender mejor el universo. Un modelo de clasificación de estrellas, galaxias y cuásares podría ayudar a los astrónomos a identificar y estudiar diferentes tipos de objetos celestes de manera más eficiente.

Identificación automática

Con un modelo de clasificación preciso, se podría automatizar la identificación de estrellas, galaxias y cuásares en grandes conjuntos de datos astronómicos. Esto ahorraría tiempo y esfuerzo en el análisis manual y permitiría un procesamiento más rápido de grandes volúmenes de datos.

Descubrimientos científicos

Un modelo de clasificación preciso podría ayudar a identificar objetos celestes raros o inusuales, lo que podría llevar a nuevos descubrimientos científicos. Por ejemplo, al identificar cuásares poco comunes, los astrónomos podrían estudiarlos con mayor detalle para comprender mejor sus propiedades y su impacto en el universo.

Aplicaciones espaciales

La clasificación precisa de objetos astronómicos es importante en el ámbito de la exploración espacial. Puede ayudar en la identificación de estrellas y galaxias de interés

para misiones espaciales específicas, como la selección de objetivos para observatorios espaciales o la planificación de rutas de sondas espaciales.

DOCUMENTACIÓN

Modelo de datos

El modelo de datos del dataset "Stellar Classification Dataset" proporcionado en Kaggle consta de las siguientes columnas:

- obj_id: Identificador único del objeto astronómico.
- ra: Ascensión recta del objeto en grados.
- dec: Declinación del objeto en grados.
- u, g, r, i, z: Magnitudes en diferentes bandas de luz (ultravioleta, verde, roja, infrarroja, infrarroja en el sistema fotométrico) utilizadas para caracterizar el brillo del objeto en diferentes longitudes de onda.
- redshift: Valor de desplazamiento al rojo basado en el aumento de la longitud de onda.
- run, rerun, camcol, field: Números de identificación asociados a la imagen y la ubicación en el campo de visión.
- spec_obj_id: Identificador único del objeto en el espectro.
- MJD: Fecha juliana modificada, utilizada para indicar cuándo se tomó una determinada pieza de datos SDSS.
- plate: Número de placa asociado a una observación astronómica específica.
- class: Clase del objeto astronómico, que puede ser "STAR" (estrella), "GALAXY" (galaxia) o "QSO" (cuásar).
- fiber_ID: Identificación de la fibra que apuntó la luz al plano focal en cada observación.

El modelo de datos tiene una fila por cada objeto astronómico en el conjunto de datos, y cada columna representa una característica específica del objeto. Las magnitudes en diferentes bandas de luz proporcionan información sobre el brillo y el color del objeto en diferentes rangos espectrales, mientras que las columnas de identificación y ubicación ayudan a rastrear y asociar los objetos con sus imágenes y posiciones en el campo de visión.

La columna *class* es la variable objetivo que se utilizará para la clasificación de estrellas, galaxias y cuásares. Dependiendo del objetivo del proyecto, se utilizará el resto de las columnas como características de entrada para el modelo de aprendizaje automático.

Para ayudar a comprender el objetivo del proyecto, se van a definir a continuación los conceptos de estrella, galaxia y cuásar.

Una **estrella** es una masa de gas caliente que emite luz y energía debido a las reacciones nucleares en su núcleo. Las estrellas se forman a partir de nubes de gas y polvo interestelar que colapsan bajo su propia gravedad. Varían en tamaño, temperatura, luminosidad y color, desde enanas rojas más pequeñas y frías hasta estrellas gigantes y supergigantes más grandes y calientes.



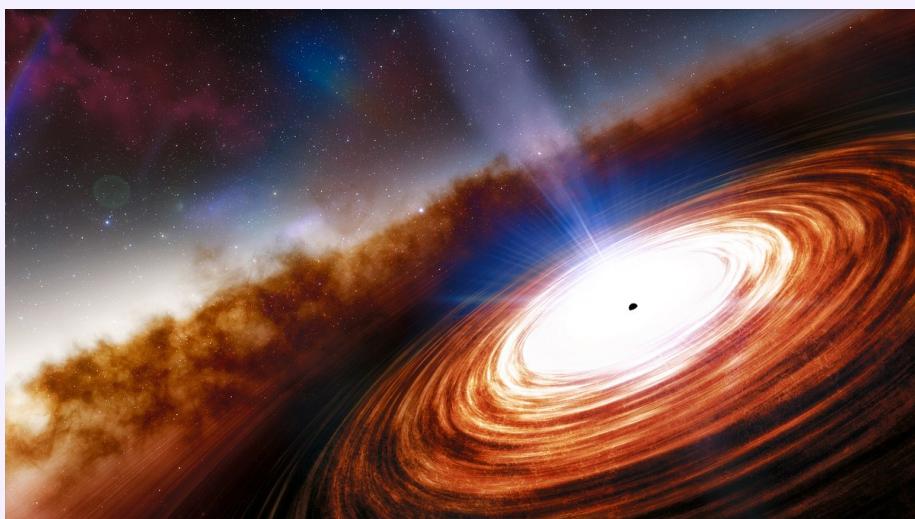
Estrellas en el espacio

Una **galaxia** es una enorme estructura cósmica compuesta por un conjunto de estrellas, planetas, gas, polvo cósmico, materia oscura y otros objetos astronómicos. Las galaxias son los principales componentes del universo y se estima que existen miles de millones de ellas en el cosmos. Pueden variar en forma y tamaño, desde galaxias en espiral con brazos distintivos hasta galaxias elípticas más suaves y galaxias irregulares. La Vía Láctea es la galaxia en la que se encuentra nuestro sistema solar.



Galaxia

Un **cuásar (quasar)** es una fuente de energía altamente luminosa y distante que se encuentra en los núcleos de galaxias lejanas. La palabra "cuásar" proviene de "fuentes de radio quasi-estelares" debido a su apariencia similar a la de una estrella en el espectro de radio. Los cuásares emiten grandes cantidades de energía en diversas frecuencias del espectro electromagnético, incluyendo radio, infrarrojo, óptico, ultravioleta y rayos X. Los cuásares son como motores cósmicos en el centro de las galaxias. Pueden ser visibles desde distancias cósmicas, lo que los convierte en objetos de estudio importantes en la cosmología y la astrofísica.



Cuásar

Definición del plan de pruebas

Las etapas de prueba ayudarán a asegurar que el modelo de clasificación estelar sea confiable, preciso y adecuado para su implementación en aplicaciones astronómicas y espaciales.

1. Prueba de integridad de datos

Verificar la calidad y la integridad del conjunto de datos utilizado para el entrenamiento y la evaluación del modelo. Esto implica asegurarse de que no haya valores faltantes, anomalías o errores en los registros.

2. Prueba de preprocessamiento de datos

Validar que el preprocessamiento de datos se haya realizado correctamente. Esto incluye verificar la normalización de características, la codificación de variables categóricas y la división adecuada del conjunto de datos en conjuntos de entrenamiento, validación y prueba.

3. Prueba de entrenamiento del modelo

Verificar que el modelo se esté entrenando correctamente y que esté convergiendo hacia una solución óptima. Esto implica monitorear las métricas de entrenamiento, como la función de pérdida y la precisión, para asegurarse de que estén mejorando de manera adecuada durante el entrenamiento.

4. Prueba de evaluación del modelo

Evaluar el rendimiento del modelo utilizando el conjunto de prueba. Calcular métricas de evaluación, como la precisión, el recall y la puntuación F1, para determinar la calidad de la clasificación. Comparar los resultados obtenidos con los objetivos establecidos para el proyecto.

5. Prueba de robustez

Realizar pruebas para evaluar la robustez del modelo frente a diferentes escenarios. Esto puede incluir la introducción de ruido en los datos de entrada o la evaluación del rendimiento del modelo en conjuntos de datos no vistos previamente.

6. Prueba de escalabilidad

Evaluar la capacidad del modelo para manejar grandes volúmenes de datos. Esto implica realizar pruebas con conjuntos de datos más grandes y verificar si el rendimiento del modelo se mantiene dentro de los límites aceptables.

7. Prueba de despliegue

Probar el modelo en un entorno de producción para verificar su rendimiento en la clasificación de nuevas observaciones de estrellas, galaxias y cuásares. Esto puede implicar el uso de datos de prueba adicionales y comparar las predicciones del modelo con las clasificaciones reales.

8. Prueba de mantenimiento y actualización

Establecer un plan para realizar pruebas periódicas y actualizar el modelo a medida que se disponga de nuevos datos o se realicen mejoras en el algoritmo. Esto garantiza que el modelo siga siendo preciso y efectivo a medida que evoluciona el entorno astronómico.

ANÁLISIS

Publicaciones del SDSS

Con cada lanzamiento de datos, la colaboración del Sloan Digital Sky Survey ha publicado un artículo de Lanzamiento de Datos, el cual describe los datos, el proceso de adquisición de datos y otros detalles del proyecto.

El lanzamiento de datos público más reciente del SDSS-IV es el Lanzamiento de Datos 17 (DR17), el cual fue lanzado en diciembre de 2021. Los detalles de DR17 se describen en el artículo del Lanzamiento de Datos 17 (Abdurro'uf et al. 2022).

Los datos publicados por el SDSS son de dominio público. El dataset está tomado de la publicación de datos RD17:

- Disponible públicamente: 6 December 2021
- Abdurro'uf et al., The Seventeenth data release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 DATA (Abdurro'uf et al. 2022 ApJS 259, 35)
- ADS abstract: [2022ApJS..259....35A](#)
- Publicación en revistas: [doi:10.3847/1538-4365/ac4414](https://doi.org/10.3847/1538-4365/ac4414)
- arXiv preprint: [arXiv:2112.02026](#)

Metodología de recolección

Versión de datos 17: imágenes, espectros ópticos (SDSS/SEGUE/BOSS/SEQUELS/eBOSS), espectros infrarrojos (APOGEE/APOGEE-2), espectros IFU (MaNGA), espectros de biblioteca estelar (MaStar) y datos de catálogo (parámetros medidos desde imágenes y espectros, como magnitudes y corrimientos al rojo).

Dimensiones de calidad del dato

Las dimensiones de calidad de datos proporcionan un marco para evaluar y mejorar la calidad de los datos en diferentes aspectos, asegurando que los datos sean confiables, precisos y adecuados para su uso en análisis y toma de decisiones.

A continuación veremos las pruebas de calidad realizadas en este conjunto de datos.

Exactitud

La exactitud se refiere a la medida en que los datos representan correctamente la realidad que intentan representar. Los datos exactos son precisos y libres de errores o inexactitudes.

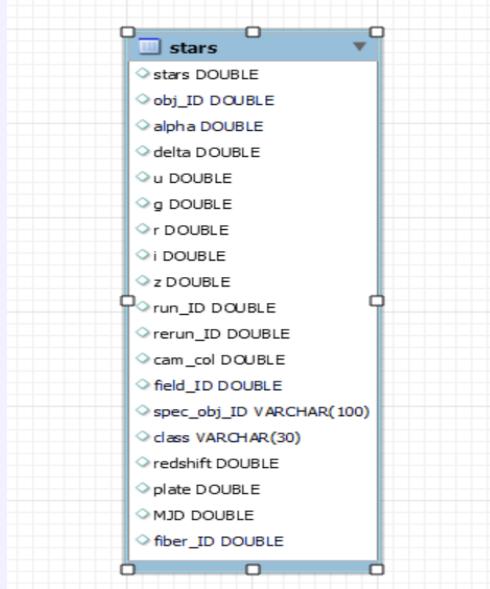
En el caso del dataset empleado en este proyecto los datos consisten en 100.000 observaciones del espacio tomadas por el SDSS (Sloan Digital Sky Survey), un proyecto de investigación espacial localizado en el observatorio Apache Point de Nuevo México. El Sloan Digital Sky Survey utiliza telescopios ópticos con una apertura angular de 2.5 metros para capturar imágenes astronómicas. Estas imágenes se adquieren utilizando un sistema fotométrico compuesto por cinco filtros distintos: u, g, r, i y z. El proceso de procesamiento de las imágenes permite obtener una lista de objetos observados, así como diversos parámetros asociados a ellos. Estos parámetros incluyen la clasificación de los objetos como puntos o de apariencia extendida, como las galaxias, y cómo el brillo registrado en los CCD se relaciona con distintas magnitudes astronómicas.

Integridad

La integridad se refiere a la consistencia y la validez de los datos. Los datos deben ser coherentes en todo el conjunto y cumplir con las restricciones y reglas establecidas para garantizar su integridad.

El dataset cumple con la regla de integridad debido a que contiene una única tabla con los IDs correctamente definidos. La tabla posee un campo de identificación único para cada registro,

asegurando que no existan duplicados o inconsistencias en los datos. Esto garantiza la integridad de los datos y facilita la realización de consultas y análisis precisos.



Unicidad

La unicidad establece que no existe más de una entidad idéntica en el mismo conjunto de datos. Es importante saber si se tiene información duplicada en formatos iguales o similares dentro de la tabla.

El dataset cumple con la regla de unicidad debido a que no contiene valores duplicados. Cada registro en el dataset es único y no se repite en ningún otro lugar. Esto garantiza que no existan datos redundantes y que cada observación sea única e individual en el conjunto de datos.

```

3. Limpieza de valores duplicados:

# Contar el número total de filas en el DataFrame
total_rows = df.count()
print(f"Filas totales en el DataFrame: {total_rows}")

# Contar el número de filas después de eliminar los duplicados
unique_rows = df.dropDuplicates().count()
print(f"Filas únicas en el DataFrame: {unique_rows}")

# Verificar si hay datos duplicados
if total_rows == unique_rows:
    print("No hay duplicados en el DataFrame.")
else:
    print(f"Hay {total_rows-unique_rows} duplicados en el DataFrame.")

Filas totales en el DataFrame: 100000
Filas únicas en el DataFrame: 100000
No hay duplicados en el DataFrame.

# Mostrar los datos duplicados
duplicated_data = df.exceptAll(df.dropDuplicates())
duplicated_data.show()

print(f"Hay {total_rows-unique_rows} filas duplicadas en el DataFrame.")

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|obj_ID|alpha|delta| u| g| r| i| z|run_ID|rerun_ID|cam_col|field_ID|spec_obj_ID|class|redshift|plate|MJD|fiber_ID|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Hay 0 filas duplicadas en el DataFrame.

```

Validez

La validez se refiere a la medida en que los datos cumplen con las reglas y restricciones definidas para su uso. Los datos válidos son legítimos y cumplen con los estándares y requisitos establecidos.

Los datos recopilados del dataset usado en este proyecto son el resultado de mediciones reales y legítimas realizadas con un telescopio, lo que garantiza su autenticidad y validez. Esto brinda confianza en la utilización de los datos para análisis y estudios relacionados con el espacio y la astronomía.

Completitud

La completitud se refiere a la medida en que todos los datos requeridos están presentes y disponibles. Los datos completos no deben tener valores faltantes y deben abarcar todos los aspectos relevantes del fenómeno o proceso que representan.

El dataset cumple con la regla de completitud puesto que no contiene valores nulos ni incompletos. Cada campo en cada registro del dataset tiene un valor asignado y no hay información faltante. Esto garantiza que todos los datos necesarios estén presentes y que no haya lagunas en la información proporcionada por el dataset.

2. Limpieza de valores nulos:

Se utiliza una expresión lambda con la función select() para aplicar sum(col(column).isNull().cast("int")) a cada columna del DataFrame. Esto cuenta el número de valores nulos en cada columna y asigna el nombre de la columna a través de alias(column). Luego utilizamos show() para mostrar el resultado.

Precisión

La precisión se refiere a la medida de exactitud y confiabilidad de los valores registrados o almacenados en un conjunto de datos. Indica la capacidad de los datos para reflejar de manera correcta y precisa la realidad que se está representando.

El dataset cumple con la regla de precisión puesto que la fuente y el origen de los datos provienen del Sloan Digital Sky Survey (SDSS), un proyecto astronómico de renombre que ha llevado a cabo una extensa recopilación de datos estelares. La reputación y el rigor científico asociados con el SDSS brindan una base sólida para confiar en la precisión de los datos.

Además, el dataset proporciona información sobre diversas características estelares, como magnitudes en diferentes bandas, colores y parámetros espectrales. Estos datos han sido recopilados utilizando tecnologías y técnicas astronómicas avanzadas, lo que favorece que haya una alta probabilidad de obtener mediciones precisas y confiables.

Consistencia

La consistencia se refiere a la uniformidad y coherencia de los datos en un conjunto de datos. Esto conlleva a que los datos sean congruentes y sigan normas predefinidas, sin presentar contradicciones lógicas ni discrepancias entre distintos atributos o registros.

El dataset cumple con la regla de consistencia puesto que los datos no contienen valores faltantes ni posibles errores en los registros. El telescopio utilizado por el SDSS lleva a cabo una serie de comprobaciones y medidas durante la adquisición de datos en tiempo real, como por ejemplo la calidad de la señal o la estabilidad instrumental.

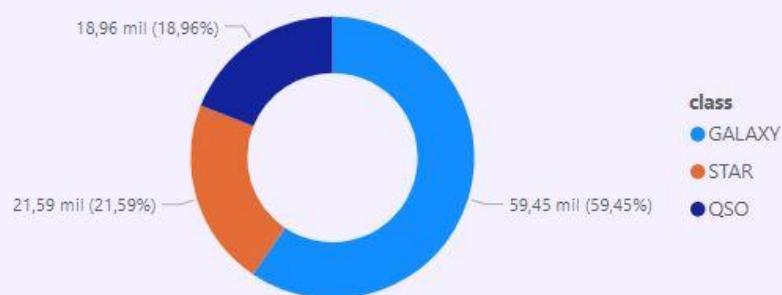
Razonabilidad

La razonabilidad se refiere a la lógica y coherencia de los datos en relación con el campo de estudio al que pertenecen los datos y al fenómeno que los datos están representando. Los datos razonables

se ajustan a las expectativas y patrones esperados en función del contexto en el que se recopilan y utilizan.

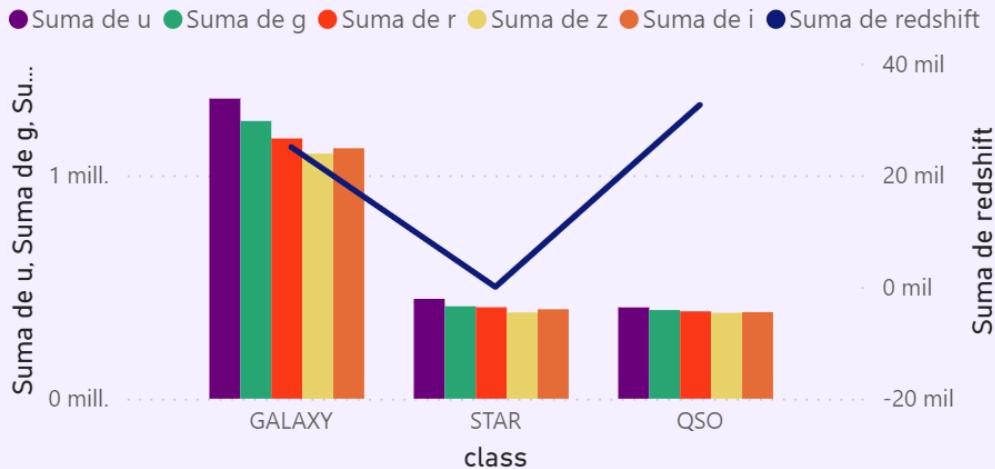
El dataset cumple con la regla de razonabilidad puesto que se garantizan la validez de los valores en relación con el campo de la astronomía y la coherencia entre los atributos. Los parámetros estelares proporcionados en el dataset se ajustan a las características conocidas de las estrellas y las propiedades establecidas por la teoría astronómica. Esto asegura que los datos son razonables en el contexto de la clasificación estelar.

Visualización

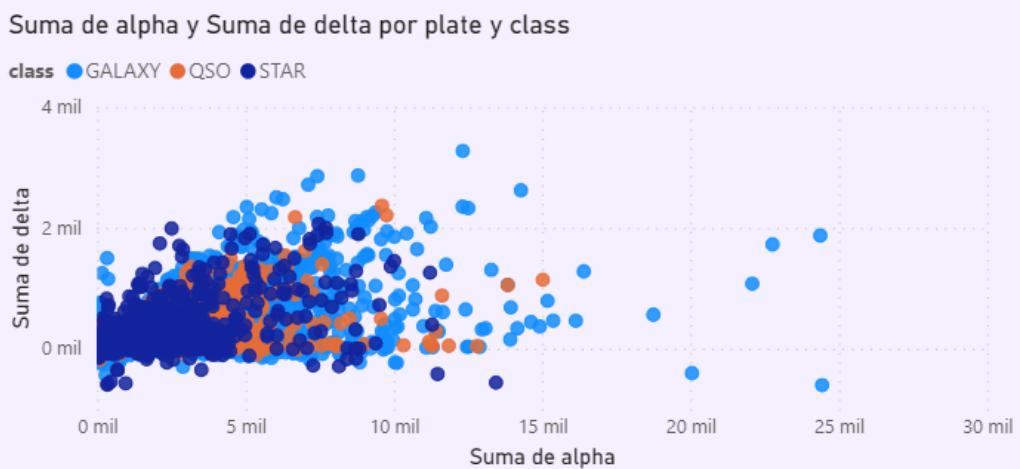


Esta gráfica muestra la distribución de estrellas, cuásares y galaxias detectadas en el dataset. Cada sección del círculo corresponde a una categoría (estrella, cuásar o galaxia) y su tamaño es proporcional a la cantidad de observaciones pertenecientes a esa categoría. De esta manera, se puede visualizar de manera intuitiva la proporción de cada tipo de objeto astronómico en el dataset y tener una idea clara de su distribución relativa.

Filtros de color y Desplazamiento al rojo(redshift) por Clases de objeto estelares



Esta gráfica muestra el desplazamiento al rojo (redshift) por cada clase de objeto estelar. En este caso, el mayor redshift se da en los cuásares, seguido de las galaxias y estrellas. Con estos datos se deduce que, a mayor desplazamiento al rojo, hay una mayor velocidad de alejamiento y posiblemente una mayor distancia con respecto al observador en la Tierra. Esto se basa en la expansión del universo, que explica que los objetos más distantes tienden a mostrar un mayor desplazamiento al rojo debido a la expansión del espacio entre ellos y la Tierra. Por esta razón los cuásares están ubicados en galaxias distantes del universo observable.



Esta gráfica de dispersión muestra la distribución de estrellas, galaxias y cuásares en el espacio, teniendo en cuenta la posición del objeto estelar en la placa del telescopio (plate) y las coordenadas angulares en el cielo (alpha y delta) para cada clase de objeto estelar. Se puede

observar que las estrellas tienen una distribución más homogénea y concentrada que los cuásares y galaxias. Las coordenadas angulares de las estrellas son menores debido a que generalmente se encuentran más cerca de la Tierra en comparación con las galaxias y los cuásares. Por otro lado, las coordenadas angulares de galaxias y cuásares son mayores en magnitud porque ocupan una porción más amplia del cielo observable y se distribuyen en áreas más extensas.

CONCLUSIÓN

En esta parte del proyecto hemos utilizado Spark y Power BI para analizar el dataset "Stellar Classification Dataset". Hemos realizado preprocesamiento de datos en Spark, pruebas de calidad y visualización en Power BI. A lo largo del proceso, nos hemos enfrentado a desafíos como la configuración de importación de datos y la manipulación de formatos. Sin embargo, hemos logrado obtener resultados precisos y demostrar la aplicabilidad de estas herramientas en el análisis de datos y la toma de decisiones basadas en información fiable.

Mediante técnicas de aprendizaje automático, se ha buscado identificar y clasificar diferentes tipos de objetos astronómicos: estrellas, cuásares y galaxias. El conjunto de datos proporcionado ha permitido entrenar y evaluar modelos de clasificación, lo que puede llegar a contribuir a mejorar nuestra comprensión del universo y a desarrollar herramientas para la identificación automática de objetos astronómicos en futuras observaciones. Este enfoque de clasificación abre nuevas posibilidades para el estudio y análisis de datos astronómicos a gran escala.