FINAL PROJECT

REBECA AMOR, PALOMA MAREQUE, LEANDRO SCANIELLO

Stellar Classification: Classification of Stars, Galaxies and Quasars

INTRODUCTION

This dataset aims to classificate stars, galaxies, and quasars based on their spectral characteristics.

Project description

The project aims to develop a machine learning model capable of classifying stars, galaxies and quasars. Using a dataset containing astronomical information, the model will be trained to distinguish between different types of celestial objects. Through the use of machine learning techniques, it seeks to achieve an accurate and efficient classification that can be applied to large astronomical data sets.

Main utilities

A model trained to classify stars, galaxies, and quasars using the dataset you mentioned can be useful in several aspects:

Astronomical research

Accurate classification of astronomical objects is crucial for better understanding the universe. A classification model for stars, galaxies, and quasars could help astronomers identify and study different types of celestial objects more efficiently.

Automatic identification

With a precise classification model, the automatic identification of stars, galaxies, and quasars in large astronomical datasets could be achieved. This would save time and effort in manually analyzing images and enable faster processing of large volumes of data.

Scientific discoveries

An accurate classification model could help identify rare or unusual celestial objects, leading to new scientific discoveries. For example, by identifying uncommon quasars, astronomers could study them in greater detail to better understand their properties and their impact on the universe.

Space applications

Precise classification of astronomical objects is important in the field of space exploration. It can aid in the identification of stars and galaxies of interest for specific space missions, such as target selection for space observatories or planning routes for space probes.

DOCUMENTATION

Data model

The data model of the "Stellar Classification Dataset" provided in Kaggle consists of the following columns:

- obj_ID: Object Identifier, the unique value that identifies the object in the image catalog used by the CAS
- alpha: Right Ascension angle (at J2000 epoch)
- delta: Declination angle (at J2000 epoch)
- u: Ultraviolet filter in the photometric system
- g: Green filter in the photometric system
- r: Red filter in the photometric system
- i: Near Infrared filter in the photometric system
- z: Infrared filter in the photometric system
- run_ID: Run Number used to identify the specific scan
- rereun_ID: Rerun Number to specify how the image was processed
- cam_col: Camera column to identify the scanline within the run
- field_ID: Field number to identify each field
- spec_obj_ID: Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec_obj_ID must share the output class)
- class: object class (galaxy, star or quasar object)
- redshift: redshift value based on the increase in wavelength
- plate: plate ID, identifies each plate in SDSS
- MJD: Modified Julian Date, used to indicate when a given piece of SDSS data was taken
- fiber_ID: fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

The data model has one row for each astronomical object in the dataset, and each column represents a specific feature of the object. The magnitudes in different light bands provide information about the brightness and color of the object in different spectral ranges, while the identification and location columns help track and associate the objects with their images and positions in the field of view.

The *class* column is the target variable that will be used for the classification of stars, galaxies, and quasars. Depending on the project's objective, the remaining columns will be used as input features for the machine learning model.

To help understand the problem, the concepts of star, galaxy and quasar will be defined below.

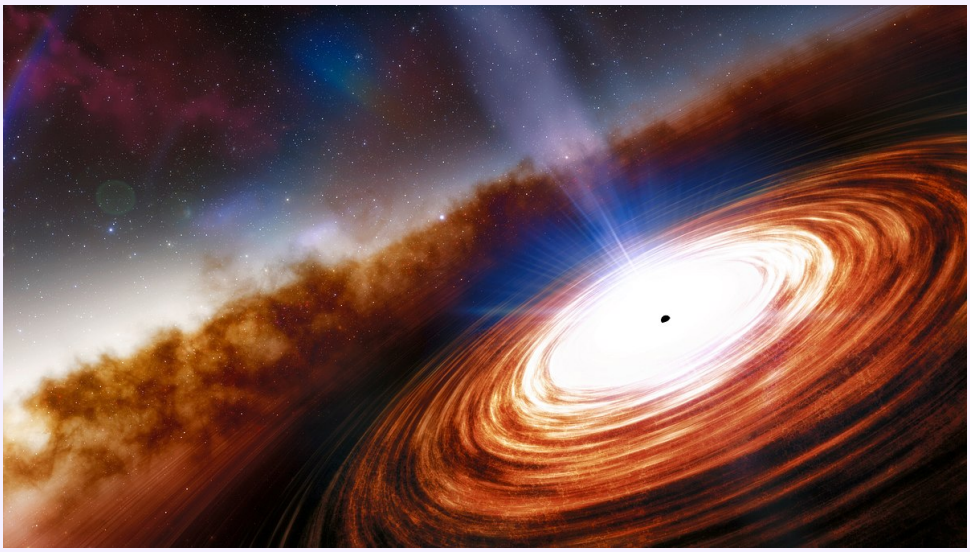
A **star** is a luminous celestial object composed of hot, glowing gases that emit light and heat. It is held together by its own gravity and generates energy through nuclear fusion. Stars vary in size, mass, and color, and they play a vital role in the universe by providing heat, light, and energy.

Stars in space

A **galaxy** is a vast system of stars, gas, dust, and other celestial objects bound together by gravity. It is one of the fundamental building blocks of the universe. Galaxies come in various shapes and sizes, ranging from small, irregularly shaped galaxies to large, spiral or elliptical galaxies.

Galaxy

A **quasar**, short for "quasi-stellar radio source," is an extremely luminous and energetic astronomical object found at the centers of some galaxies. Quasars emit enormous amounts of energy, particularly in the form of electromagnetic radiation, including visible light, ultraviolet light, and X-rays.



Quasar

Test plan

The testing stages will help ensure that the stellar classification model is reliable, accurate, and suitable for implementation in astronomical and space applications.

1. Data integrity test

Verify the quality and integrity of the dataset used for model training and evaluation. This involves ensuring that there are no missing values, anomalies, or errors in the records.

2. Data preprocessing test

Validate that the data preprocessing has been performed correctly. This includes verifying feature normalization, categorical variable encoding, and proper splitting of the dataset into training, validation, and test sets.

3. Model training test

Verify that the model is being trained correctly and converging towards an optimal solution. This involves monitoring training metrics such as loss function and accuracy to ensure they are improving appropriately during training.

4. Model evaluation test

Evaluate the performance of the model using the test set. Calculate evaluation metrics such as accuracy, recall, and F1 score to determine the quality of classification. Compare the obtained results with the project's objectives.

5. Robustness test

Perform tests to assess the model's robustness against different scenarios. This may include introducing noise in the input data or evaluating the model's performance on previously unseen datasets.

6. Scalability test

Evaluate the model's ability to handle large volumes of data. This involves testing with larger datasets and verifying if the model's performance remains within acceptable limits.

7. Deployment test

Test the model in a production environment to verify its performance in classifying new observations of stars, galaxies, and quasars. This may involve using additional test data and comparing the model's predictions with actual classifications.

8. Maintenance and update test

Establish a plan for conducting periodic tests and updating the model as new data becomes available or algorithm improvements are made. This ensures that the model remains accurate and effective as the astronomical environment evolves.

DATASET ANALYSIS

SDSS Data Release Publications

The data released by the SDSS is under public domain. It's taken from the current data release RD17:

- Publicly Available: 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
- Journal publication: doi:10.3847/1538-4365/ac4414
- arXiv preprint: arXiv:2112.02026

Collection methodology

Data Release 17 - images, optical spectra (SDSS/SEGUE/BOSS/SEQUELS/eBOSS), infrared spectra (APOGEE/APOGEE-2), IFU spectra (MaNGA), stellar library spectra (MaStar), and catalog data (parameters measured from images and spectra, such as magnitudes and redshifts).

Data quality dimensions

The data quality dimensions provide a framework for evaluating and improving data quality in different aspects, ensuring that the data is reliable, accurate, and suitable for use in analysis and decision-making.

Next we will see the quality tests performed on this dataset.

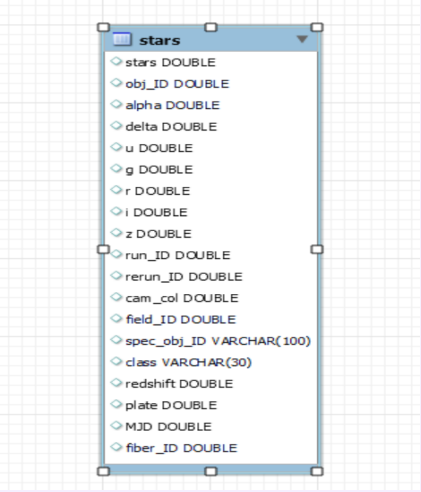
Precision

Precision refers to the extent to which data correctly represents the reality they are intended to represent. Exact data is accurate and free from errors or inaccuracies.

In the case of the dataset used in this project, the data consists of 100,000 space observations taken by the Sloan Digital Sky Survey (SDSS), a space research project located at the Apache Point Observatory in New Mexico. The Sloan Digital Sky Survey utilizes optical telescopes with an angular aperture of 2.5 meters to capture astronomical images. These images are acquired using a photometric system composed of five different filters: u, g, r, i, and z. The image processing process allows for the extraction of a list of observed objects, along with various parameters associated with them. These parameters include the classification of objects as point-like or extended in appearance, such as galaxies, and how the brightness recorded in the CCD relates to different astronomical magnitudes.

Integrity

Integrity refers to the consistency and validity of the data. The data must be consistent throughout the set and comply with the restrictions and rules established to guarantee its integrity.



The dataset complies with the integrity rule as it consists of a single table with well-defined IDs. The table includes a unique identification field for each record, ensuring there are no duplicates or inconsistencies in the data. This ensures data integrity and facilitates accurate querying and analysis.

Uniqueness

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.

The dataset complies with the uniqueness rule because it does not have any duplicate values. Each record in the dataset is unique and does not repeat anywhere else. This ensures that there is no redundant data and that each observation is unique and distinct within the dataset.

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.
```

Validity

Validity refers to the extent to which the data complies with the rules and restrictions defined for its use. Valid data is legitimate and meets established standards and requirements.

The data represents real and legitimate measurements taken with a telescope, ensuring its authenticity and validity. This instills confidence in using the data for analysis and studies related to space and astronomy.

Completeness

Completeness refers to the extent to which all required data is present and available. Complete data must not have missing values and must cover all relevant aspects of the phenomenon or process they represent.

The dataset complies with the completeness rule as it does not have any null or incomplete values. Each field in every record of the dataset has an assigned value, and there is no missing information. This ensures that all necessary data is present and there are no gaps in the information provided by the 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.

```
from pyspark.sql.functions import col, sum

# Obtener el número total de nulos por columna
null_counts = df.select([sum(col(column).isNull().cast("int")).alias(column) for column in df.columns])
null_counts.show()
```

```
+------+-----+-----+---+---+---+---+---+------+--------+-------+--------+-----------+-----+--------+-----+---+--------+
|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|
+------+-----+-----+---+---+---+---+---+------+--------+-------+--------+-----------+-----+--------+-----+---+--------+
|     0|    0|    0|  0|  0|  0|  0|  0|     0|       0|      0|       0|          0|    0|       0|    0|  0|       0|
+------+-----+-----+---+---+---+---+---+------+--------+-------+--------+-----------+-----+--------+-----+---+--------+
```

Accuracy

Accuracy refers to the measure of accuracy and reliability of the values recorded or stored in a data set. It indicates the ability of the data to correctly and accurately reflect the reality or phenomenon being represented.

The dataset complies with the completeness rule since the source and origin of the data are from the Sloan Digital Sky Survey (SDSS), a renowned astronomical project that has conducted extensive stellar data collection. The reputation and scientific rigor associated with the SDSS provide a solid basis for confidence in the accuracy of the data.

In addition, the dataset provides information on various stellar characteristics, such as magnitudes in different bands, colors, and spectral parameters. These data have been collected using advanced astronomical technologies and techniques, which increases the likelihood of obtaining accurate and reliable measurements.

Consistency

Consistency refers to the uniformity and coherence of data in a dataset. This means that the data is consistent and follows predefined rules and conventions, without presenting logical contradictions or discrepancies between different attributes or records.

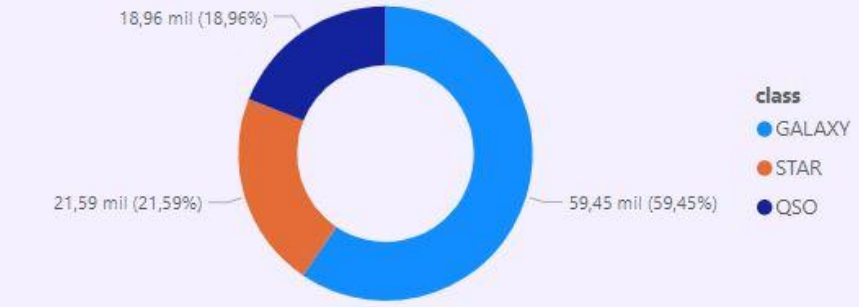
The dataset meets the consistency rule as it does not contain missing values or potential errors in the records. This is ensured by the telescope used by the SDSS, which performs a series of checks and measurements during real-time data acquisition, such as signal quality or instrumental stability.

Reasonability

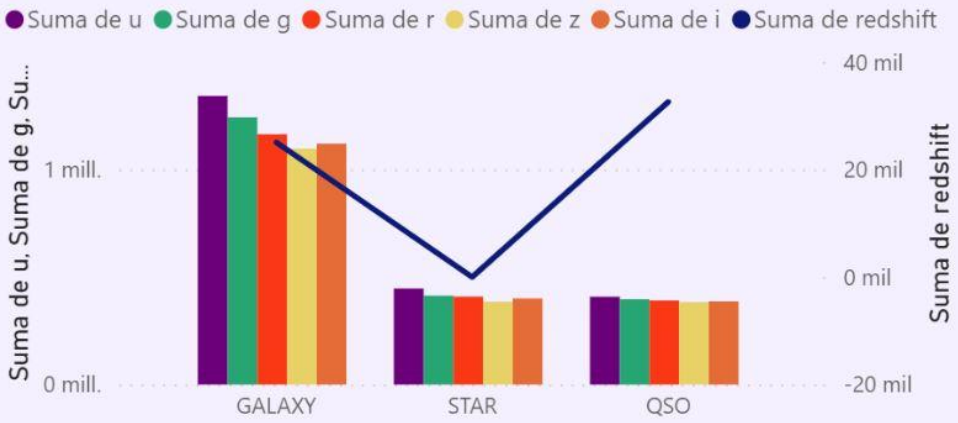
Reasonability refers to the logic and coherence of data in relation to the field of study to which the data belongs and the phenomenon the data represents. Reasonable data aligns with expectations and expected patterns based on the context in which they are collected and used.

The dataset complies with the reasonability rule as it ensures the validity of values in relation to the field of astronomy and coherence among attributes. The stellar parameters provided in the dataset align with the known characteristics of stars and the properties established by astronomical theory. This ensures that the data is reasonable within the context of stellar classification.

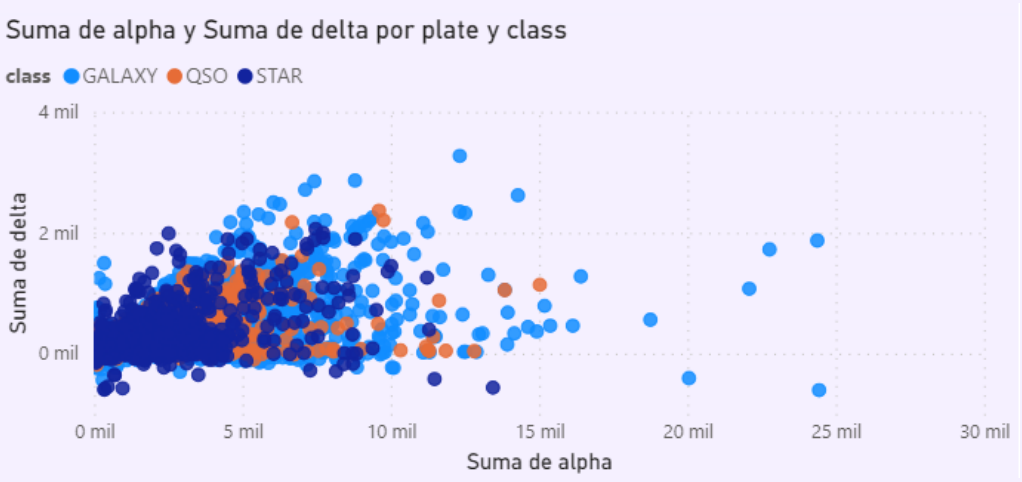
Data visualization



This graph shows the distribution of stars, quasars, and galaxies detected in the dataset. Each section of the circle corresponds to a category (star, quasar or galaxy) and its size is proportional to the number of observations belonging to that category. In this way, you can intuitively visualize the proportion of each type of astronomical object in the dataset and have a clear idea of their relative distribution.



This graph shows the redshift for each stellar object class. In this case, the largest redshift occurs in quasars, followed by galaxies and stars. With these data it can be deduced that the greater the redshift, there is a greater speed of departure and possibly a greater distance with respect to the observer on Earth. This is based on the expansion of the universe, which explains that more distant objects tend to show a greater redshift due to the expansion of space between them and Earth. For this reason quasars are located in galaxies distant from the observable universe.



This scatter plot shows the distribution of stars, galaxies, and quasars in space, taking into account the position of the celestial object on the telescope plate ("plate") and the angular coordinates in the sky ("alpha" and "delta") for each class of celestial object. It can be observed that stars have a more homogeneous and concentrated distribution compared to quasars and galaxies. The angular coordinates of stars are smaller because they are generally closer to Earth compared to galaxies and quasars. On the other hand, the angular coordinates of galaxies and

quasars are larger in magnitude because they occupy a wider portion of the observable sky and are distributed over more extensive areas.

CONCLUSION

In this part of the project, we have used Spark and Power BI to analyze the "Stellar Classification Dataset". We have performed data preprocessing in Spark, quality testing, and visualization in Power BI. Throughout the process, we encountered challenges such as data import configuration and format manipulation. However, we have successfully obtained accurate results and demonstrated the applicability of these tools in data analysis and making informed decisions based on reliable information.

Using machine learning techniques, we have sought to identify and classify different types of astronomical objects: stars, quasars and galaxies. The provided data set has allowed us to train and evaluate classification models, which may contribute to improve our understanding of the universe and to develop tools for the automatic identification of astronomical objects in future observations. This classification approach opens up new possibilities for the study and analysis of astronomical data on a large scale.