# SDSS Stellar Classification

Kirsten Bauck

## Introduction

I chose to explore this stellar classification dataset because of my fascination with astronomy and the challenges it presents in physics and data science. The dataset, derived from the Sloan Digital Sky Survey (SDSS), tackles one of astronomy's fundamental tasks: classifying celestial objects. In this case, classifying them as stars, galaxies, or quasars. This can be a challenge when it comes to identifying minority classes like quasars, which despite being less common, play a much needed role in helping us understand our universe.

This data is significant as it plays a foundation role in astronomical research. Stellar classification is key to understanding the fundamental properties of celestial objects, including their temperature, composition, and evolutionary stage. The more scientists can accurately classify stellar objects, the better they can understand how such objects interact with each other. Classifications such as these have historically led to discoveries such as the distinction between the Andromeda galaxy and our own Milky Way. From this, I hope to learn how data science techniques can be used to analyze and explore a dataset in such a way that makes future modeling more efficient.

Regarding the data source, I have a low level of confidence. While this dataset is available on Kaggle and was curated by a data scientist with a double masters credentials, it represents a processed version of SDSS data. This curated process, while probably making the datasource more accessible for analysis does add a layer of separation from a primary source and should not be used in a real analysis

## Data Cleaning

Since this data was curated, the data cleaning process was relatively simple and few challenges were encountered. The data cleaning process worked as follows and was done in R. First, I calculated and removed any extreme outliers from the dataset. Then I changed the modified Julian Date (MJD) column to a simple date column with year, month, day. Finally, I removed any unnecessary columns within the dataset, such as extraneous ID fields and renamed columns to be more easily interpretable.

The dataset consists of 99,999 observations collected by SDSS, describing celestial objects on 10 features. The key variables are split into the 3 following main types and sample data is shown in Table 1.

1. Classification label & Date
    - **Class** consists of three categories GALAXY (majority class), QSO (quasars), and STAR
    - **Date** date object was observed by SDSS
2. Positional data
    - **Right Ascension angle** which is the astronomical equivalent of longitude
    - **Declination angle** which is the astronomical equicalent of latitude
3. Spectral measurements:
    - **Ultraviolet filter**, **Green filter**, **Red Filter**, **Near Infrared Filter** and **Infrared Filter** readings
    - **Redshift** measurements (indicating the Doppler shift of light)

Table 1: Sample of SDSS Stellar Object Classification Data

| Class | Right Ascension Angle | Declination Angle | Ultraviolet Filter | Green filter | Red Filter | Near Infrared Filter | Infrared Filter | Redshift | Date |
|---|---|---|---|---|---|---|---|---|---|
| GALAXY | 135.69 | 32.49 | 23.88 | 22.28 | 20.40 | 19.17 | 18.79 | 0.63 | 2013-03-03 |
| GALAXY | 144.83 | 31.27 | 24.78 | 22.83 | 22.58 | 21.17 | 21.61 | 0.78 | 2018-02-09 |
| GALAXY | 142.19 | 35.58 | 25.26 | 22.66 | 20.61 | 19.35 | 18.95 | 0.64 | 2011-01-31 |
| GALAXY | 338.74 | -0.40 | 22.14 | 23.78 | 21.61 | 20.50 | 19.25 | 0.93 | 2017-10-13 |
| GALAXY | 345.28 | 21.18 | 19.44 | 17.58 | 16.50 | 15.98 | 15.54 | 0.12 | 2012-09-17 |
| QSO | 341.00 | 20.59 | 23.49 | 23.34 | 21.32 | 20.26 | 19.55 | 1.42 | 2011-10-21 |
| QSO | 23.23 | 11.42 | 21.47 | 21.18 | 20.93 | 20.61 | 20.43 | 0.59 | 2018-12-04 |
| GALAXY | 5.43 | 12.07 | 22.25 | 22.02 | 20.34 | 19.49 | 18.85 | 0.48 | 2012-10-10 |
| GALAXY | 200.29 | 47.20 | 24.40 | 22.36 | 20.61 | 19.46 | 18.96 | 0.66 | 2013-04-04 |
| STAR | 39.15 | 28.10 | 21.75 | 20.03 | 19.18 | 18.82 | 18.65 | 0.00 | 2006-12-13 |
| GALAXY | 328.09 | 18.22 | 25.77 | 22.52 | 20.64 | 19.78 | 19.06 | 0.46 | 2011-10-18 |
| GALAXY | 243.99 | 25.74 | 23.77 | 23.80 | 20.98 | 19.81 | 19.46 | 0.59 | 2011-05-12 |
| STAR | 345.80 | 32.67 | 23.17 | 20.14 | 19.42 | 19.22 | 18.89 | 0.00 | 2013-09-05 |
| GALAXY | 331.50 | 10.04 | 20.83 | 18.75 | 17.51 | 17.02 | 16.63 | 0.15 | 2011-06-27 |
| GALAXY | 344.98 | -0.35 | 23.21 | 22.79 | 22.09 | 21.86 | 21.85 | 0.82 | 2016-10-21 |
| GALAXY | 244.82 | 25.15 | 24.89 | 22.13 | 20.45 | 19.49 | 18.97 | 0.48 | 2011-05-12 |
| STAR | 353.20 | 3.08 | 24.55 | 21.44 | 20.95 | 20.79 | 20.48 | 0.00 | 2011-10-30 |
| QSO | 1.49 | 3.29 | 20.39 | 20.41 | 20.30 | 20.06 | 19.89 | 2.03 | 2016-10-02 |
| STAR | 14.38 | 3.21 | 21.82 | 20.56 | 19.95 | 19.76 | 19.56 | 0.00 | 2015-12-17 |
| GALAXY | 167.13 | 67.34 | 20.48 | 18.68 | 17.62 | 17.12 | 16.73 | 0.11 | 2001-01-20 |
| GALAXY | 171.98 | 67.75 | 22.13 | 20.85 | 18.97 | 18.32 | 17.98 | 0.37 | 2001-02-02 |
| STAR | 144.79 | 46.83 | 24.55 | 22.34 | 20.92 | 19.87 | 19.17 | 0.00 | 2014-02-22 |
| GALAXY | 145.27 | 46.96 | 25.44 | 20.77 | 19.66 | 19.08 | 18.83 | 0.66 | 2002-02-11 |
| QSO | 145.88 | 47.30 | 21.74 | 21.53 | 21.27 | 21.36 | 21.16 | 2.08 | 2014-02-21 |
| GALAXY | 241.43 | 27.22 | 18.88 | 17.54 | 17.02 | 16.75 | 16.72 | 0.03 | 2003-07-04 |
| GALAXY | 132.92 | 4.52 | 21.26 | 20.50 | 18.36 | 23.18 | 17.96 | 0.25 | 2003-01-31 |
| GALAXY | 143.29 | 5.55 | 25.98 | 21.31 | 19.61 | 18.83 | 18.28 | 0.46 | 2012-01-02 |
| GALAXY | 140.60 | 5.27 | 25.47 | 22.41 | 21.43 | 20.26 | 19.99 | 0.61 | 2011-12-28 |
| GALAXY | 333.31 | -0.38 | 20.53 | 18.84 | 18.05 | 17.60 | 17.29 | 0.09 | 2003-06-23 |
| GALAXY | 337.09 | -0.31 | 20.15 | 18.37 | 17.31 | 16.82 | 16.44 | 0.15 | 2001-08-22 |

## Data Visualizations

The distribution of redshifts in Figure 1 provides a clear view of the spatial arrangement of stellar objects in our observable universe. Stars, with their redshift values clustered at zero, are primarily local objects within the Milky Way galaxy. Redshift is a measure of the stretching of light waves as objects move away from us, with higher redshifts corresponding to greater distances. The galaxies, in contrast, exhibit a broader distribution of redshifts between 0 and 1, reflecting their presence in the nearby and mid-distance universe. Quasars, identifiable by their high redshifts (up to 3.5), are among the most distant objects visible to us. Their extreme luminosity, powered by supermassive black holes, allows them to be visible at great distances where normal stars or galaxies would be too faint.

This visualization shows the distinct clustering of different stellar object types in our universe, with stars mostly within the local universe, galaxies spread through nearby regions, and quasars giving us a glimpse into the early cosmos.

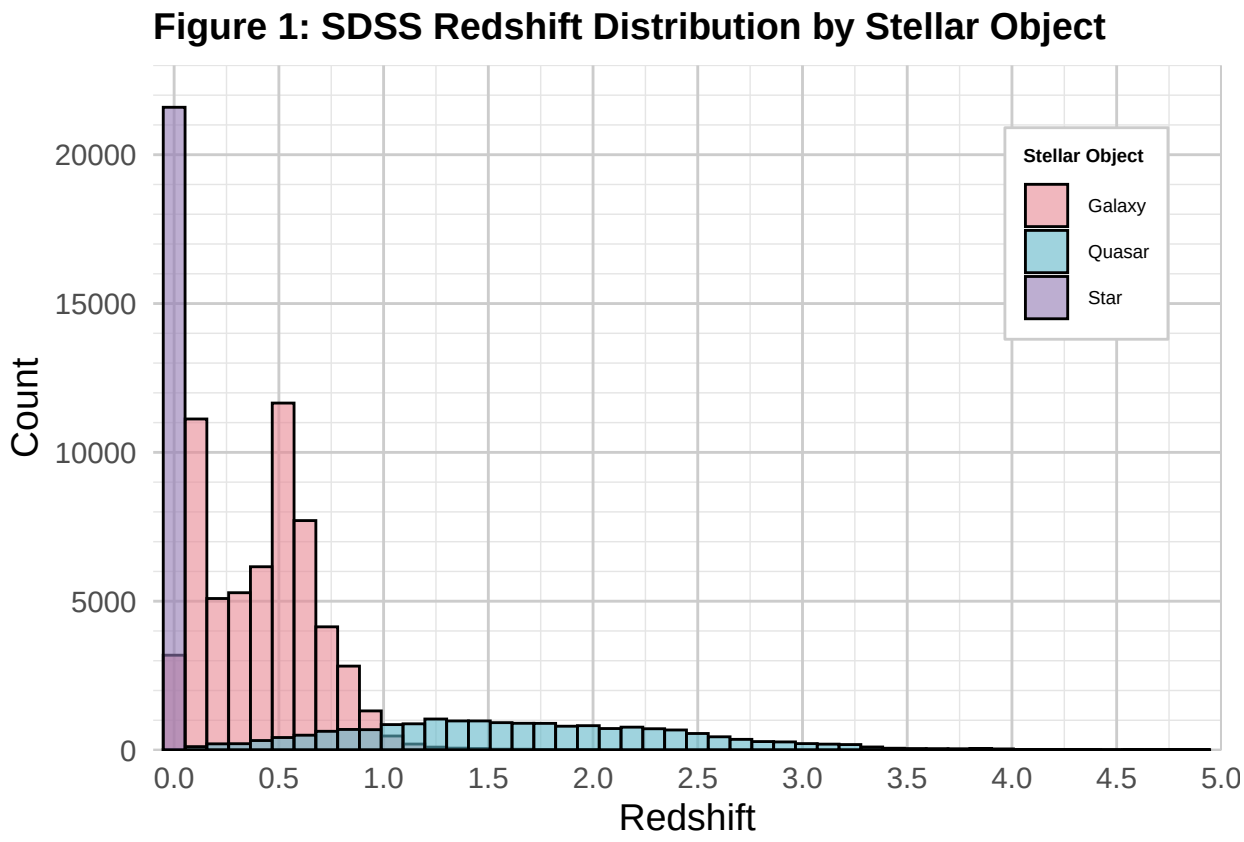


*Figure 1: Distribution of redshifts across stellar objects (stars, galaxies, and quasars) in the SDSS dataset, highlighting the different distances of these objects from Earth.*

Figure 2 maps the right ascension (RA) and declination (Dec) of stellar objects in the northern sky. The plot is a polar projection, placing familiar constellations and notable objects like the Big Dipper, Andromeda Galaxy, and Polaris within a wider cosmic context. The dataset shows galaxies as the dominant object type, evenly distributed across the surveyed regions, with quasars scattered more sparsely but still following similar patterns. Stars are also present, forming an arc-like structure between the two main survey regions.

By avoiding the dusty galactic plane, the SDSS focuses on clearer skies that enhance the detection and classification of extragalactic objects. The inclusion of famous constellations and objects helps orient the viewer, offering a familiar backdrop for understanding the cosmic distribution of these stellar objects.

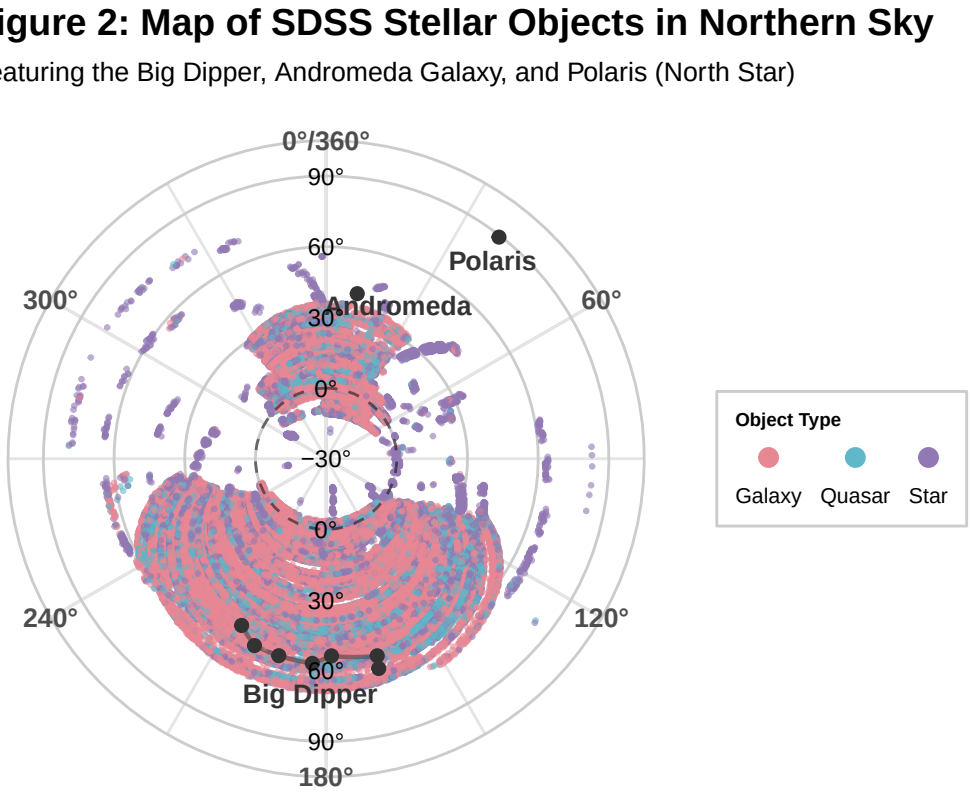


*Figure 2: Polar map of SDSS stellar objects in the northern sky, showing their distribution relative to the Big Dipper, Polaris, and other key cosmic features.*

## Conclusions

Given more time, there are a couple more analysis that I would have done. First of all, I would want to do a full feature importance analysis, where I could quantify the contribution of each feature to a model. In addition, I would also like to use techniques such as PCA to visualize the class separability in the high dimensional space of all the variables (filters, redshift, and position) interacting. Finally, I would also want to go more in depth into an error analysis, where I examine misclassified objects and try to identify any systemic biases or challenging regions in the space.

In a previous analysis of this SDSS data, various machine learning models were applied to classify the stellar objects. A summary of the models, their performance, advantages, and limitations is provided in Table 2.

Table 2: Model Summary and Comparison

| Model | Test Error | Advantages | Limitations |
|---|---|---|---|
| Decision Tree | 4.17 | Interpretable, Visualizable | Prone to overfitting, Unstable |
| Random Forest | 2.06 | Handles non-linear relationships, Robust to outliers | Computationally expensive, Less interpretable |
| AdaBoost | 2.41 | Reduces bias and variance, Focuses on misclassified examples | Sensitive to outliers, Computationally intensive |

The Random Forest model, having an ensemble apprpoach, demonstrated the lowest test error, indicating its effectiveness at capturing some of the complex patters in the dataset while also reducing noise. AdaBoost also performed well, as it was able to leverage its ability to refine predictions by focusing on misclassified examples but it was computationaly expensive. Throughout my analysis, it seemed that the redshift variable was the most important in classifying the stellar objects. From this previous analysis, I learned that balancing interpretability (Decision Tree) and accuracy (Random Forest) is a key challenge, especially as a datasets' dimensions increase and modeling becomes more computationaly expensive. Before my analysis I did not fully understand the point of an intepretable model that was known to perform less well. But after conducing my analysis, I now understand that the more interpretable models are necessary to try and understand what less interpretable models are doing in the background.

When it comes to analyzing this dataset further, there are some potential challenges. First of all, the class imbalance of the dataset consisting mostly of galaxies can easily bias any model created. In addition, as one tries to incorporate additional spectral features, analysis could become more computationally intensive and harder to understand. The fact that most of the filters are highly correlated could lead to redundancy in models, which then complicates feature importance interpretation and could even affect model stability. Most of this data is photometric as well, which is less precise compared to spectroscopic data, so the reliance on such data to classify objects could limit the classification accuracy, especially when distinguish between quasars and stars.

## Link to GitHub Repository

**Click here to install from GitHub**

## Citations

fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved November 30th, 2024 from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.