IOSR Journal of Applied Physics (IOSR-JAP)
e-ISSN: 2278-4861.Volume 17, Issue 1 Ser. 2 (Jan. – Feb. 2025), PP 51-58
www.iosrjournals.org

Comparative Analysis for Celestial Classification Using Machine Learning

Adhyatma Chulet

Abstract
The objective of this paper is to perform star/galaxy/QSO classification using machine learning techniques and the spectroscopic data from the Sloan Digital Sky Survey (SDSS). In particular, the dataset contains many labelled spectral classes of stars (G, K, M, F, A), which will be utilised for stellar classification. Photometric features in multiple bands, including u, g, r, i, z, J, and H, serve as input features for the models. For star/galaxy/QSO classification, machine learning algorithms such as k-nearest neighbour (KNN), random forest (RF), and support vector machine (SVM) exhibit strong performance. For stellar classification, RF and SVM provided an advantage over KNN, and RF dominated all other models. Individually, RF provided a maximum accuracy of 97.17%, SVM provided an accuracy of 93.70%, and KNN was 63.66%. The area under the receiver operating characteristic (ROC) curve for all four models approaches 1, showing high classification reliability. Additionally, evaluation metrics such as accuracy, precision, recall, F-score, and Matthews correlation coefficient consistently exceed 0.5. This broad range of metrics led to an accurate and reliable assessment of model performance. The confusion matrices also indicated that the highest percentage of erroneous classifications occurred between galaxies and QSOs. Overall, these models effectively predict the nature of sources and accurately classify stellar types in the SDSS spectroscopic dataset.

I. Introduction

Astronomy has evolved from simple pictures of the night sky to large, massive digital movies of space in motion. Modern telescopes today are now collecting huge volumes of data every day, with some astronomical instruments expected to produce 750 terabytes of data each second once the Australian Square Kilometre Array Pathfinder is fully operational by 2025. However, with the volume of astronomical data available today, there are many opportunities to apply sophisticated computational techniques and algorithms to analyse that data to better understand our universe.

Stars, galaxies and quasars are celestial objects that can be categorised based on their many varying properties. A star is any massive self-luminous celestial body of gas that shines by radiation derived from its internal energy sources [9]. Galaxies are vast assemblages of stars, gas, dust and dark matter, held together by gravity and appearing in various shapes like spirals, ellipses or irregular clusters [5]. Quasars, or quasi-stellar objects, are extremely bright active galactic nuclei (AGNs) that are powered by a supermassive black hole.

The Sloan Digital Sky Survey (SDSS) has played an important role in advancing our current understanding of celestial objects. It combines high-resolution imaging with spectroscopic data, and its detailed database contains numerous pieces of data that can be useful in classification, including redshifts, compositions, declinations, and motions.

By categorising photometric and spectral data about these astronomical objects into different classes, we can gain valuable insights into their underlying physical processes and properties and how they can be valuable for predicting future astronomical phenomena.

Before major advancements in technology, this task was performed using labour-intensive manual analysis of spectral data, a process that, while effective, isn't able to match the rate at which the available datasets grow. Machine learning allows us to process large amounts of data, making this process immensely easier [3,4]. Multiple studies have shown the extreme potential of machine-learning algorithms such as Support Vector Machines (SVM) and Random Forests in classifying stars and galaxies [5].

However, classification using these machine learning algorithms is subject to some problems as well. Some of these algorithms have difficulty distinguishing between different features of astronomical bodies due to their extreme similarity, and this process also requires vast amounts of computational power to analyse and compare the data. Therefore, it is imperative to determine which algorithms have the highest success rate in classifying these celestial bodies.

II. Literature Review

The classification of celestial bodies has been an important research area for many years. Scientists have employed various methods and machine learning (ML) models to improve how accurately we can classify stars, galaxies, and quasars. Earlier studies focused on traditional statistical models, whereas newer approaches use advanced Machine Learning (ML) techniques to achieve better results.

B. Arsioli and P. Dedin[2] studied Support Vector Machines (SVMs) and other ML algorithms to classify blazars based on their synchrotron peak frequencies and multifrequency features. They found that SVMs had a balanced accuracy of 93%, with spectral slopes and infrared colours being key factors in classification.

Sunitha P. et al. [8] applied Multi-Class SVM (MSVM) to classify astronomical objects using image data. Their method involved preprocessing images with Otsu's thresholding and extracting spatial and texture features. MSVM proved to be highly efficient in distinguishing stars, galaxies, and extended objects, making it useful for image-based classification.

D. Fraix-Burnet et al. [6] explored unsupervised classification using the Fisher-EM algorithm on over 700,000 galaxy spectra from the SDSS. This method grouped galaxies into 86 spectral classes with a 15% error rate. Unlike traditional clustering methods, Fisher-EM handled high-dimensional data better by focusing on the most important features, making it particularly useful for large-scale galaxy classification.

Yulun Wu [5] emphasised the need for large sample sizes to improve classification accuracy. The study compared Logistic Regression and Decision Trees for classifying stars, galaxies, and quasars using SDSS data. The Decision Tree model achieved 99% accuracy, performing much better than Logistic Regression's 87%. Decision Trees' ability to handle imbalanced datasets made them a strong choice for classification.

Yu Bai et al. [1] tested random forests to categorise stars, galaxies, and quasars. Their study showed that random forests could achieve over 99% accuracy for stars and galaxies and 94% accuracy for quasars using data from the SDSS and the Large Sky Area Multi-Object Fibre Spectroscopic Telescope. Random forests handled overlapping features and imbalanced datasets better than traditional methods while keeping low processing costs.

Zhichen Lin [4] used K-Nearest Neighbours (KNN) combined with Principal Component Analysis (PCA) to classify stars, galaxies, and quasars. PCA reduced unnecessary features, making the model more efficient. The KNN model achieved an accuracy of 96%-98%, proving that selecting the right features can improve model performance. Lin also noted that feature scaling and dimensionality reduction techniques, such as the Standard Scaler and PCA, can reduce problems caused by dominant features and improve classification accuracy.

Tracy X. Chen et al. [3] applied ML to a real-world problem by integrating it into the NASA/IPAC Extragalactic Database (NED) for classifying astrophysical journal articles. Their model had over 90% accuracy in replicating human classifications, showing how ML can efficiently handle large amounts of data in astronomy.

Although ML models have many advantages, they also face some challenges. Performance can be affected by class imbalances and feature relationships. Wu [5] noted that datasets with fewer quasars could lead to biased predictions. Decision Trees address class imbalance by using impurity-based splitting criteria. Lin [4] emphasised that even similar stellar objects can have variations in their features, making large and high-quality training datasets essential.

Many studies highlight the importance of selecting the appropriate algorithm for specific datasets. Arsioli and Dedin[2] found that SVMs worked best for multifrequency spectral data, while Wu [5] showed that Decision Trees performed well for photometric datasets. Lin [4] demonstrated that combining KNN with PCA improved efficiency, proving that different datasets require different approaches.

Overall, advances in ML have significantly improved celestial classification, with newer methods continuously refining earlier models. As datasets grow larger and more complex, further research into hybrid models and automated feature selection will be essential to improving classification techniques.

III. Methodology

This paper implements data from the Sloan Digital Sky Survey (SDSS), a comprehensive astronomical survey that provides detailed imaging and mapping of the universe [5]. The SDSS collected extensive amounts of multi-wavelength data using a 2.5-meter telescope that have made significant contributions to cosmology, galaxy formation, and stellar evolution [10]. Our paper particularly uses the most recent data set, Data Release 18, part of the SDSS-V project. Our study incorporates several variables from the SDSS dataset to differentiate between stars, galaxies and QSOs, as outlined in Table 1.

Table 1: Variable Descriptions

U	The brightness of the object in the ultraviolet (U) band corresponds to a wavelength of around 3551 Å.
G	The brightness of the object in the green (G) band was measured at a wavelength near 4686 Å.
R	The brightness of the object in the red (R) band represents light with a wavelength close to 6166 Å.
I	The brightness of the object in the near-infrared (I) band, with a wavelength of approximately 7480 Å.
Z	The brightness of the object in the far-infrared (Z) band is measured at a wavelength around 8932 Å.
Redshift	A measure of how far and fast the object is moving away from Earth.
Declination(dec)	The angular position of the object relative to the celestial equator.
Class	The type of celestial object, such as a star, galaxy, or quasar.

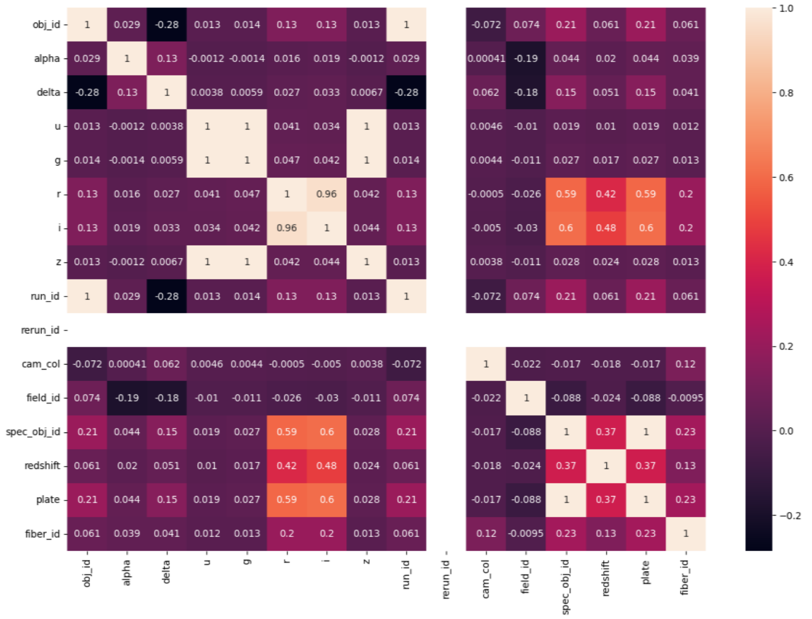

Figure 3.1: Correlation Matrix Heat Map

The correlation matrix, shown in Fig. 3.1, elucidates the link between various factors, such as redshift, delta, run_id, rerun_id, and Target. For instance, a factor like redshift may correlate with an object's apparent brightness/luminosity or type, as shown when looking at quasars, which often produce higher redshifts and distinct brightness characteristics due to them being located at greater distances.

Our study mainly focuses on redshift and declination as the primary features for the classification of stars, galaxies and quasars. This is due to the fact that redshift, which indicates an object's distance and velocity, is majorly different for each celestial object, as portrayed in Fig. 3.2: stars typically are subject to low redshift, galaxies generate moderate redshift, and quasars display high redshift, demonstrating their immense distance from our planet. Moreover, declination aids us in discerning between spatial distribution patterns of stars, galaxies and quasars due to it being a key indicator of the precise location of astronomical bodies in space.

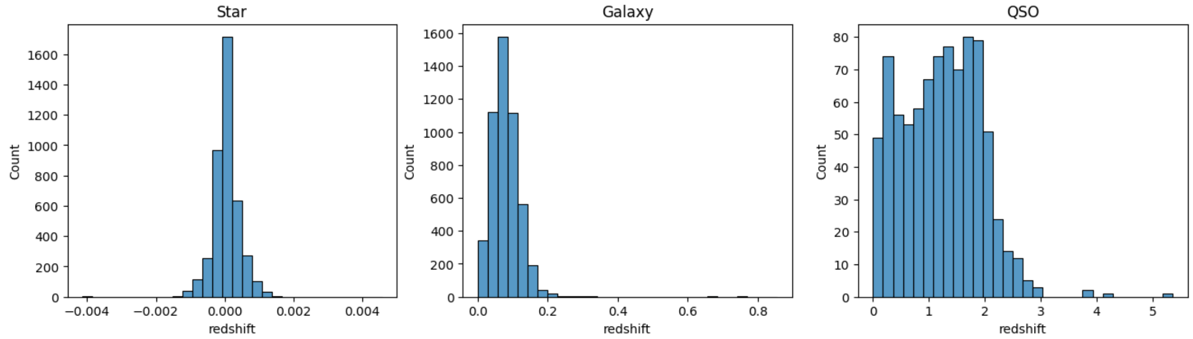

Figure 3.2

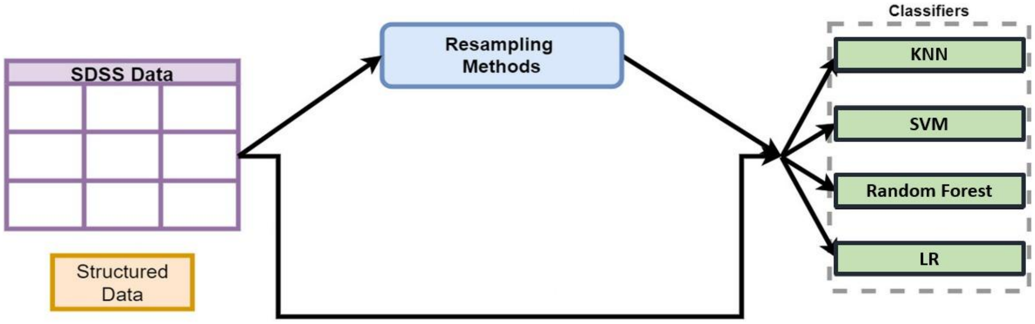

Figure 3.3

Figure 3.3 displays the process followed in this paper. It begins with the input of structured SDSS data into software. Then, the data is processed using various resampling methods to address inconsistencies, such as class imbalances and to normalise feature values, ensuring that each celestial object is represented equally. We made use of some machine learning methods like K-Nearest Neighbours (KNN), Random Forest (RF), and Support Vector Machines (SVMs) to analyse and sort info about space objects and their trajectories.

KNN determines the category of a given data point by evaluating the categories of its nearest neighbours based on a distance metric, such as Euclidean distance. The number of neighbours (K) plays a crucial role in the model's performance, where a smaller K value can lead to noise sensitivity, and a larger K value can cause over-smoothing.

Given a dataset with S number of samples, each observation belongs to a class Y_i, and the classification of a new observation X' is determined using equation 1:

$$Y' = \frac{1}{K}\sum_{i=1}^{K} Y_i$$

where the K nearest neighbours are chosen based on the following distance formula:

$$d\left(X_i, X_j\right) = \sqrt{\sum_{k=1}^{n}\left(X_{ik} - X_{jk}\right)^2}$$

where X_{ik} and X_{jk} are feature values for samples i and j, respectively.

KNN is highly dependent on feature scaling, as features with larger magnitudes dominate the distance calculation. Therefore, normalisation using a Standard scalar is crucial for improving KNN's performance. It works well for grouping stars or sorting galaxies, but it can slow down when there's a lot of data to compare.

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions for improved accuracy and robustness. It is widely used for classification and regression due to its ability to handle large datasets and reduce overfitting. The algorithm operates by creating a collection of decision trees where each tree is trained on a bootstrap sample of the dataset, and a random subset of features is considered for splitting at each node.

Given a dataset with S number of samples, each observation X_i belongs to a class Y_i, and the final prediction of Random Forest for a new observation X' is obtained using equation 6:

$$Y' = \text{mode}\left(T_1\left(X'\right), T_2\left(X'\right), \ldots, T_m\left(X'\right)\right)$$

where $T_1, T_2, \ldots, T_m$ represent individual decision trees in the ensemble.

Each tree makes an independent prediction, and the final classification is determined by majority voting in classification tasks or averaging in regression tasks. The randomness introduced by bootstrapping and feature selection helps Random Forest mitigate overfitting and improve generalization. Additionally, Random Forest can efficiently handle datasets with missing values and categorical features.

SVMs are good at drawing clear lines between data categories, which helps sort objects like stars and planets. They work well with small or unclear datasets, which is particularly handy in space missions with limited or imperfect data.

IV. Results

The study evaluated how tree depth and the number of trees affect the accuracy of a Random Forest model. Deeper trees consistently improved accuracy, but the improvements became negligible beyond a certain point. Also, more trees improved performance, but the improvement was less noticeable after 200 trees. The max accuracy of 97.17% was obtained with a depth of 8 and 200 trees; however, similar results were seen with other large configurations. Overall, accuracy can be improved by adding more trees and increasing the depth of the trees.

The K-Nearest NeighFigure 4.2 is a confusion matrix that shows how well the KNN classification model distinguishes between galaxies (class 0), QSOs (class 1), and stars (class 2). The model has achieved an accuracy rate of 97.18%, indicating that it's doing a great job overall, especially with stars, as it correctly classified all of them.

However, there were some errors with 130 galaxies being classified as QSOs, and 170 QSOs being identified as galaxies. This mix-up suggests that galaxies and QSOs have some similar features, which we discussed in this paper. Additionally, there was one minor error where a QSO was misclassified as a star. These errors indicate that we need to distinguish features between galaxies and QSOs better to improve accuracy

bours (KNN) algorithm was evaluated for its performance in classifying galaxies, quasars (QSOs), and stars. The results indicate that the accuracy of the algorithm improves as the number of neighbours (k) increases. At k=1, the accuracy was 58.81%, reflecting the algorithm's reliance on a single neighbour, which can lead to errors in noisy or overlapping datasets. With increasing values of k, the accuracy improved consistently, with the maximum recorded improvement of 63.66% when k=10.

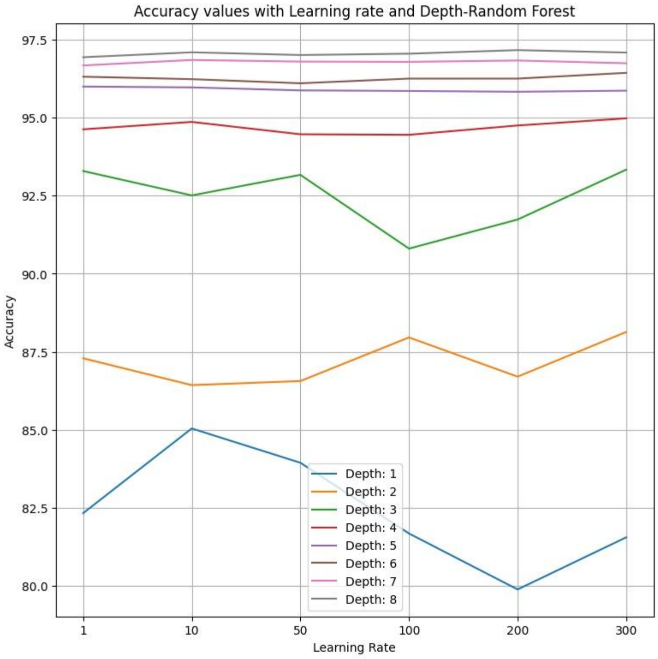

Figure 4.1: Accuracy versus Number of Trees in Random Forest

Figure 4.1 displays an Accuracy graph when the number of trees and depth change. As the depth and number of trees increase, the accuracy of the Random Forest model also increases. Deeper trees (depth 6-9) achieve the highest accuracy (97-98%), while shallower trees (depth ≤3) don't perform well, especially as tree count increases. An optimal balance of Depth 6-9 with 100-200 trees ensures high accuracy and efficiency.

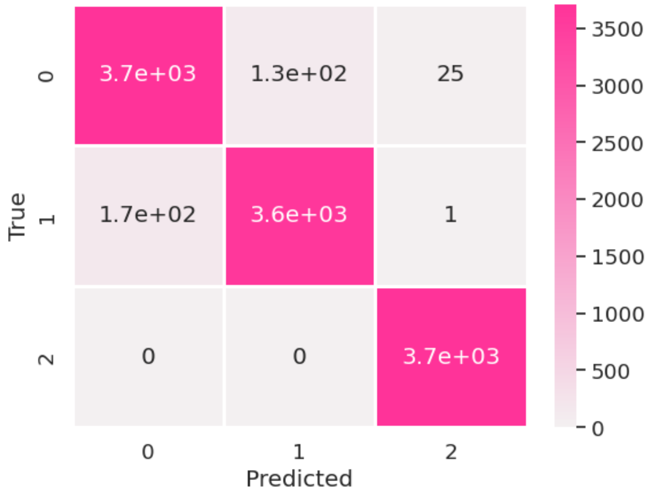

Figure 4.2 Confusion Matrix for KNN Classification

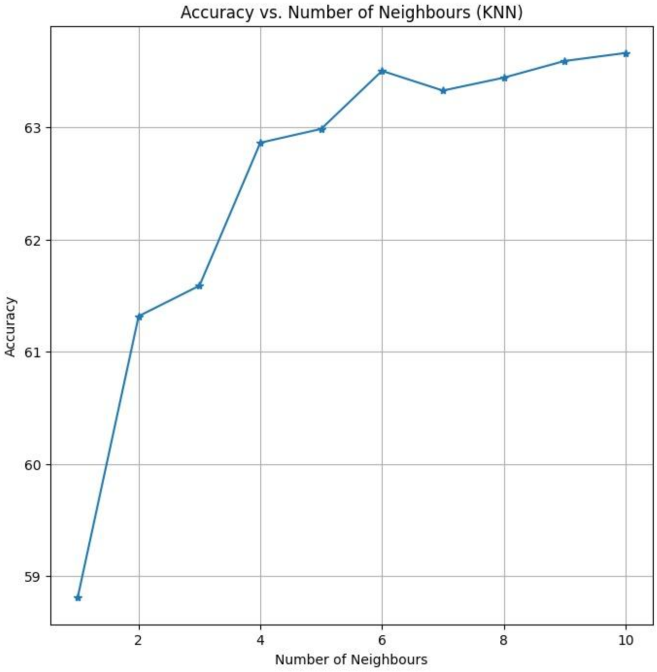

Figure 4.3 KNN Accuracy vs. Number of Neighbours

Figure 4.3 demonstrates the relationship between the KNN algorithm's classification accuracy and the number of nearest neighbours(k). Generally, as k increases, the accuracy of the KNN model improves. Initially, lower k values result in substandard performance due to sensitivity to noise, while larger k values stabilise accuracy. This trend in the values of k also suggests that regulating k is crucial for optimizing KNN, as excessively small or large values can impact classification effectiveness.

V. Conclusion

In conclusion, this study analysed the classification of stars, quasars (QSOs), and galaxies using redshift and declination as key features, applying Random Forest (RF), K-nearest neighbours (KNN), and Support Vector Machine (SVM) models. Among them, Random Forest demonstrated the highest accuracy of 97.17%, showing its exemplary ability to handle complex astronomical data and capture complex decision boundaries. SVM followed next with an accuracy of 93.7%, proving its effectiveness in distinguishing celestial objects, particularly in high-dimensional feature spaces. Lastly, KNN achieved an accuracy of 63.66%, portraying its limitations due to sensitivity to distance metrics and data distribution.

The findings presented above highlight the advantages of ensemble-based and kernel-driven methods and algorithms in astronomical classification. The superior performance of RF and SVM indicates that employing additional features or using advanced, modern techniques such as deep learning could further enhance accuracy. Future research can also delve into the impact of feature engineering, feature selection and larger datasets to further refine model performance. Advancements in machine learning algorithms, which could allow them to perform classification directly using images/visual data, will also produce higher accuracy. Ultimately, implementing machine learning for celestial categorisation has already had multiple significant benefits for sky surveys and large-scale astronomical research.

References

[1]. Bai, Yu, et al. "Machine learning applied to Star–Galaxy–QSO classification and stellar effective temperature regression." The Astronomical Journal 157.1 (2018): 9.

[2]. Arsioli, Bruno, and Pedro Dedin. "Machine learning applied to multifrequency data in astrophysics: blazar classification." Monthly Notices of the Royal Astronomical Society 498.2 (2020): 1750-1764.

[3]. Chen, Tracy X., et al. "Classification of Astrophysics Journal Articles with Machine Learning to Identify Data for NED." Publications of the Astronomical Society of the Pacific 134.1031 (2022): 014501.

[4]. Lin, Zhichen. "Classification of GALAXY, QSO, and STAR Based on KNN and PCA."

[5]. Wu, Yulun Winston. "MACHINE LEARNING CLASSIFICATION OF STARS, GALAXIES, AND QUASARS."

[6]. Fraix-Burnet, Didier, Charles Bouveyron, and J. Moultaka. "Unsupervised classification of SDSS galaxy spectra." Astronomy & Astrophysics 649 (2021): A53.

[7]. Xiao-Qing, Wen, and Yang Jin-Meng. "Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches." Chinese Journal of Physics 69 (2021): 303-311.

[8]. Sunitha, P., Naveen Venkatesh, and M. Atreya. "Support vector machine method to identify and classify astronomical objects." International Journal of Engineering Applied Sciences and Technology 5 (2020): 209-215.

[9]. Fernie, John Donald, Aller, Lawrence Hugh, Brecher, Kenneth, Chaisson, and Eric J. 2025. “Star | Definition, Light, Names, & Facts.” Encyclopedia Britannica. April 2, 2025. https://www.britannica.com/science/star-astronomy.

[10]. “Sloan Digital Sky Survey - DR18.” 2023. Kaggle. July 29, 2023. https://www.kaggle.com/datasets/diraf0/sloan-digital-sky-survey-dr18/data.