Stellar Classification Using Spectral Analysis

Manas R. Pandya - IITM Zanzibar

February 14, 2024

Abstract

Stellar classification is a fundamental task in understanding the universe. With the advent of large-scale surveys, automated classification using machine learning is quite helpful.

In this report, I have presented an approach to stellar classification of galaxies, stars and quasars using spectral analysis with the latest data from the Sloan Digital Sky Survey Data Release 17 (SDSS DR17) (clickable link) .

1. Introduction

I have applied a machine learning algorithm to automate the classification of stellar objects (galaxies, stars and quasars) using their spectral data. **The k-Nearest Neighbors (kNN) algorithm, known for its simplicity and effectiveness, is chosen for this task.** I have trained the model to accurately predict the class of a celestial object based on its spectral characteristics derived from the SDSS DR17.
The Sloan Digital Sky Survey (SDSS) represents one of the most ambitious and influential surveys in the history of astronomy. Its seventeenth data release, SDSS DR17, is where the dataset for my project obtained from.

1.1. Dataset Description

The dataset utilized in this study encompasses a diverse array of spectral measurements of over 100,000 stellar objects. The measurements include the stellar object's Redshift, the magnitudes in u,g,r,i, and z filters
Apart from these, there is other data that can be used for more complicated machine learning techniques. These include the location co-ordinates of the stellar object in the night sky, temperature and other IDs of spectral data.
Visual excerpt of the dataset is presented below in Figure 1.

obj_ID	class	redshift	u	g	r	i	z	alpha	delta
1237660961327	GALAXY	0.6347936	23.87882	22.2753	20.39501	19.16573	18.79371	135.6891066	32.49463184
1237664879951	GALAXY	0.779136	24.77759	22.83188	22.58444	21.16812	21.61427	144.8261006	31.27418489
1237660961330	GALAXY	0.6441945	25.26307	22.66389	20.60976	19.34857	18.94827	142.1887896	35.58244418
1237663478724	GALAXY	0.9323456	22.13682	23.77656	21.61162	20.50454	19.2501	338.7410378	-0.4028275746
1237680272041	GALAXY	0.1161227	19.43718	17.58028	16.49747	15.97711	15.54461	345.2825932	21.1838656
1237680272039	QSO	1.424659	23.48827	23.33776	21.32195	20.25615	19.54544	340.9951205	20.58947628
1237678858481	QSO	0.5864546	21.46973	21.17624	20.92829	20.60826	20.42573	23.23492643	11.41818762
1237678858473	GALAXY	0.477009	22.24979	22.02172	20.34126	19.48794	18.84999	5.433176037	12.06518599
1237661435386	GALAXY	0.660012	24.40286	22.35669	20.61032	19.4649	18.95852	200.2904754	47.19940232
1237670961088	STAR	-7.90E-06	21.74669	20.03493	19.17553	18.81823	18.65422	39.1496906	28.10284161
1237680272034	GALAXY	0.4595958	25.77163	22.52042	20.63884	19.78071	19.05765	328.0920762	18.22031048
1237662341088	GALAXY	0.5914091	23.76761	23.79969	20.98318	19.80745	19.45579	243.9866375	25.73828043
1237680507721	STAR	7.18E-05	23.17274	20.14496	19.41948	19.22034	18.89359	345.8018744	32.67286785
1237678858459	GALAXY	0.1521936	20.8294	18.75091	17.51118	17.01631	16.62772	331.50203	10.03580205
1237663478726	GALAXY	0.8181597	23.20911	22.79291	22.08589	21.86282	21.8512	344.9847703	-0.3526157812
1237662341088	GALAXY	0.4849288	24.8868	22.13311	20.44728	19.49171	18.9747	244.8245231	25.15456399
1237678598087	STAR	-0.000428576	24.5489	21.44267	20.95315	20.7936	20.48442	353.2015224	3.080795936
1237678598091	QSO	2.031528	20.38562	20.40514	20.29996	20.05918	19.89044	1.494388639	3.29174633
1237678598096	STAR	-0.0004402762	21.82154	20.5573	19.94918	19.76057	19.55514	14.38313522	3.214326196
1237651539783	GALAXY	0.1115879	20.48292	18.67807	17.6168	17.11936	16.73351	167.1316688	67.33993563
1237651539783	GALAXY	0.3747563	22.13367	20.84772	18.96537	18.31696	17.98124	171.9754246	67.74745014

Figure 1: Snippet of the stellar classification dataset.

2. Details about the Data and Feature Selection

The SDSS photo metric system uses five filters designated as u, g, r, i, and z, each measuring the intensity of light from celestial objects at specific wavelengths.

- **u filter:** Captures near-ultraviolet light, centered around 355 nm.
- **g filter:** Measures light in the green part of the spectrum, centered around 477 nm.
- **r filter:** Sensitive to red light, centered around 623 nm.
- **i filter:** Covers the near-infrared range, centered around 763 nm.
- **z filter:** Extends into the infrared, centered around 913 nm.

The magnitude measured in each filter is a logarithmic scale that quantifies the brightness of an object as seen from Earth.

2.1. Understanding Redshift

Redshift refers to the phenomenon where light from an object is shifted towards the red end of the spectrum as the object moves away from us. This effect is a result of the Doppler shift, analogous to the change in pitch of a siren as an ambulance drives by. In the cosmic sense, Redshift is used to determine the velocity at which objects such as galaxies are receding from Earth, providing critical information about the expansion rate of the universe and the distance to these objects.

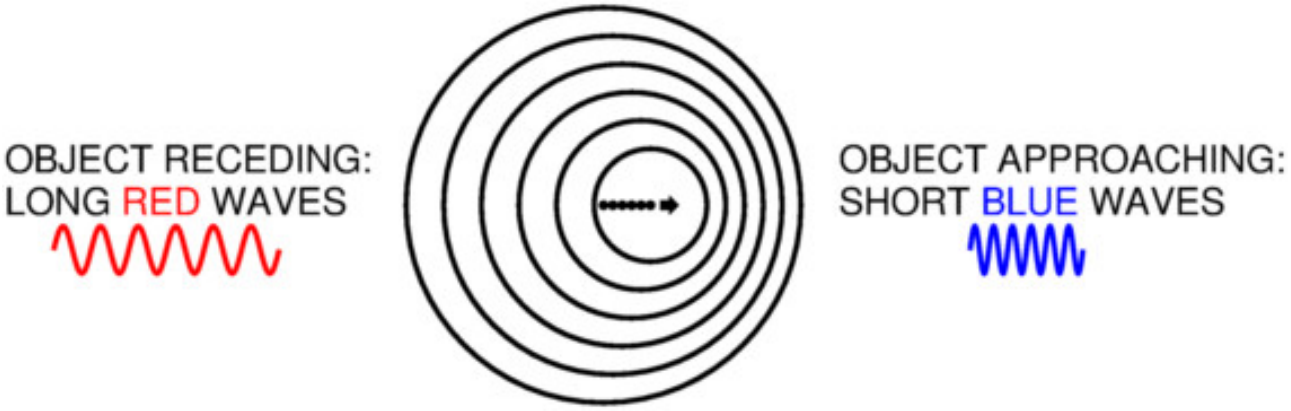


Figure 2: Redshift visualisation

2.2. Color Indices and Their Importance

In spectral analysis, the color indices, which are the differences between the magnitudes measured in different filters (e.g., u-g, g-r), are **more informative than the magnitudes** themselves. This is because color indices effectively represent the color of the object, which can reveal properties such as temperature, age, and metal content.
For instance, a high g-r index might indicate a cooler, older star, while a lower index suggests a hotter, younger star. This can be extrapolated to galaxies and quasars, helping us classify them.

2.3. Justification for Feature Selection in kNN Model

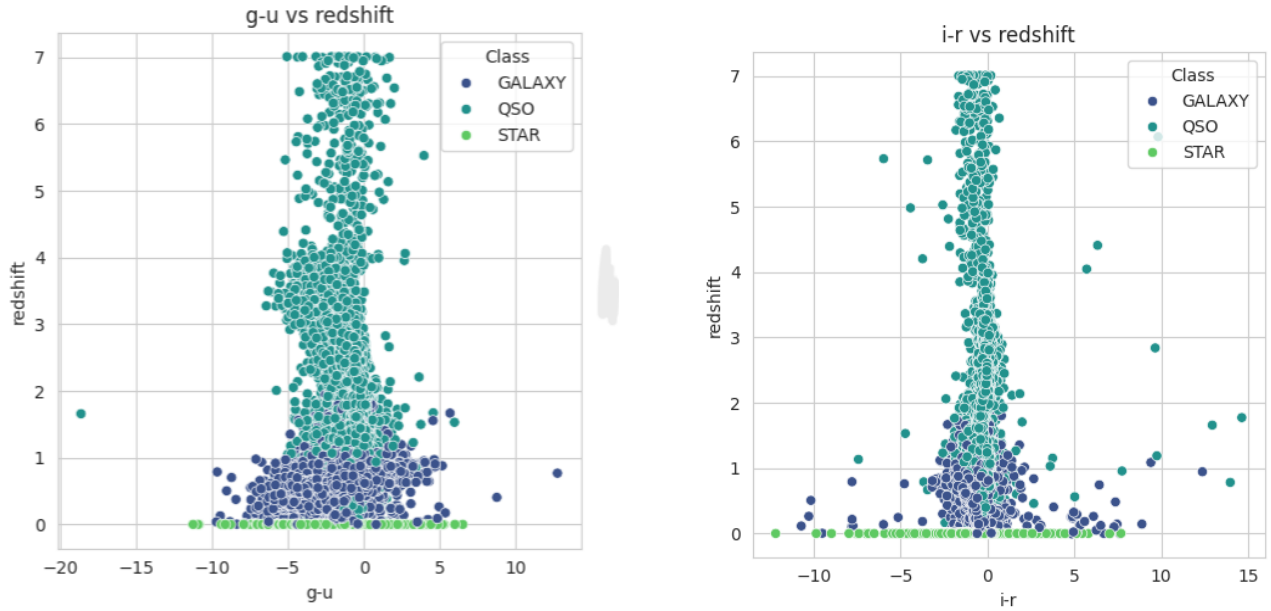
When it comes to classifying celestial objects using machine learning models like kNN, color indices are particularly valuable. They are less sensitive to distance and extinction effects than absolute magnitudes. Since kNN relies on the similarity between instances, using color indices as features ensures that the model is comparing intrinsic properties of objects rather than their apparent brightness, which can be affected by their distance and intervening dust.
Therefore, the difference in filter magnitudes, or color indices, are used as features in the kNN model to enhance the classification performance and provide a more accurate analysis of the inherent characteristics of the objects in the dataset.

2.4. Visual proof for feature selection

Initial analysis visualized the relationships between spectral features, color indices, for classification. These indices provide more meaningful data for classification than individual magnitudes, which can be affected by brightness and distance.

Distinct clusters in the plots of g-u vs Redshift and i-r vs Redshift allowed for clear separation between stars, galaxies, and quasars, making them suitable for kNN classification. Figures 3a and 3b demonstrate their advantage.

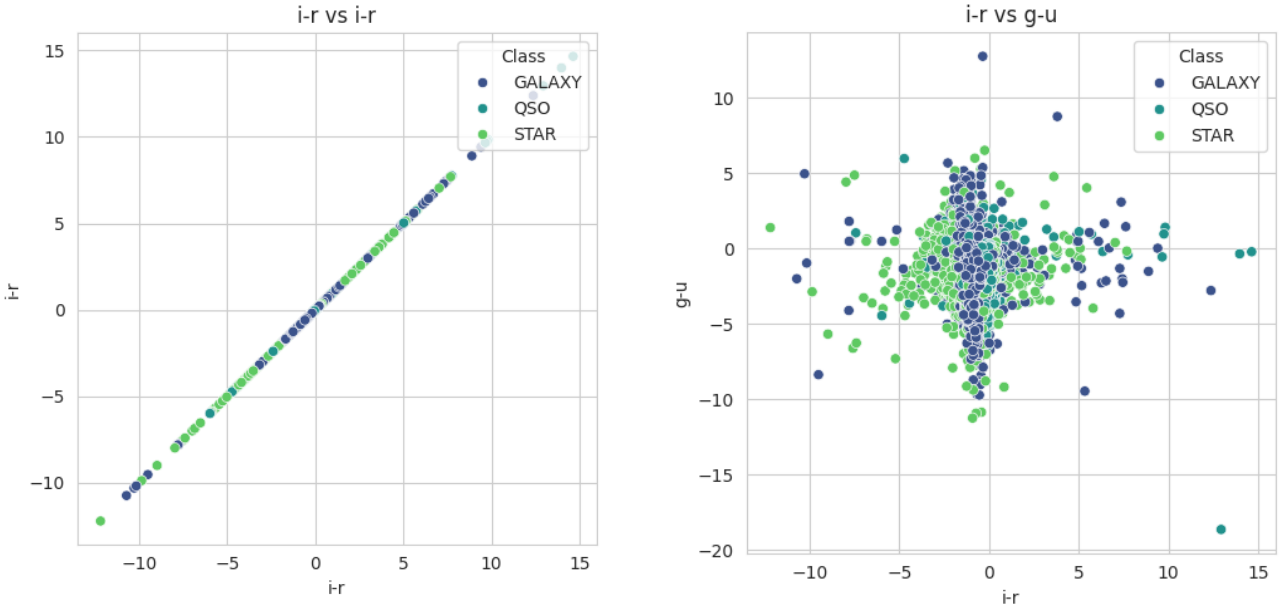
In contrast, other combinations like i-r vs i-r and i-r vs g-u did not provide clear class separations, as seen in Figure 4. The i-r vs i-r plot is redundant, while i-r vs g-u shows significant class overlap, making them unsuitable for kNN classification. .



(a) g-u vs Redshift

(b) i-r vs Redshift

Figure 3: Scatter plots showing **clear class separations.**



(a) i-r vs i-r

(b) i-r vs g-u

Figure 4: Scatter plots showing **inadequate class separation.**

The chosen features, g-u vs Redshift and i-r vs Redshift, offer higher accuracy and better insights into the spectral characteristics of celestial objects, justifying their use for further analysis.

3. Methodology

3.1. k-Nearest Neighbors (kNN) Algorithm

The k-Nearest Neighbors algorithm is a simple, yet effective machine learning technique used for classification tasks. **It operates on the principle that similar data points are often in close proximity.**
This is visualised accurately in the Figure 5. The algorithm trains on a subset of data to cluster the different classes based upon their features magnitudes (the axes of the graph).
For a new data point, the model will calculate the distance between the clusters, and classify the data point to a particular class based upon it's proximity. My model can be visualised as like given in the figure, with the features on the axes (u-g and i-r on the x axis and Redshift on the y axis).
However, kNN can be implemented for Multidimensional data as well, i.e, data with more than two features. But of course that couldn't be visualised on a graph. Hence I have proceeded with two features for the current project.

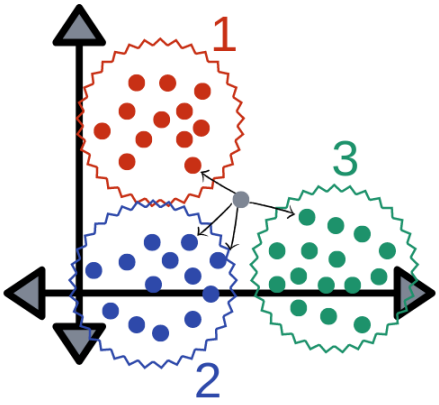


Figure 5: kNN visualisation

3.2. Training and Testing the Model

Training the Model refers to the process of teaching the algorithm to make predictions or decisions, based on data. It involves providing the algorithm with input features - color indices (u-g, i-r) and Redshift and the corresponding target values - classes: star, galaxy, quasar. During training, the algorithm learns to associate patterns in the input features with the target values. In the case of kNN, the training process involves storing the features and target values.
This is because kNN is a lazy learner :) that doesn't build a model in the same way other algorithms do, but rather memorizes the training dataset.
In my project 80% of the dataset (around 81,000 stellar objects' data) has been used for training - whilst the rest 20,000 have been used for testing the model

Testing the Model is the phase where the trained model is evaluated to see how well it performs on unseen data. It involves providing a new set of features (the test set) to the model and asking it to predict the target values. The predictions are then compared to the actual target values of the test set. The outcome of this comparison, often in terms of accuracy, and other metrics, provides a measure of the model's performance. Testing helps to assess the model's ability to generalize to new, unseen data, beyond the examples it was trained on.
In my project the metrics used to evaluate the model are accuracy and the confusion matrix (this is a set of True and false Predictions).

3.3. Results

Our kNN classification model was evaluated using a test dataset to determine its accuracy and overall performance in classifying celestial objects as stars, galaxies, or quasars. The accuracy of the model provides a straightforward metric for assessing its effectiveness.

3.3.1. Accuracy:

The model achieved an accuracy of **94.48% for the g-u vs redshift** classification and **95.62% for the i-r vs redshift** classification. These values indicate the proportion of total predictions that were correct.

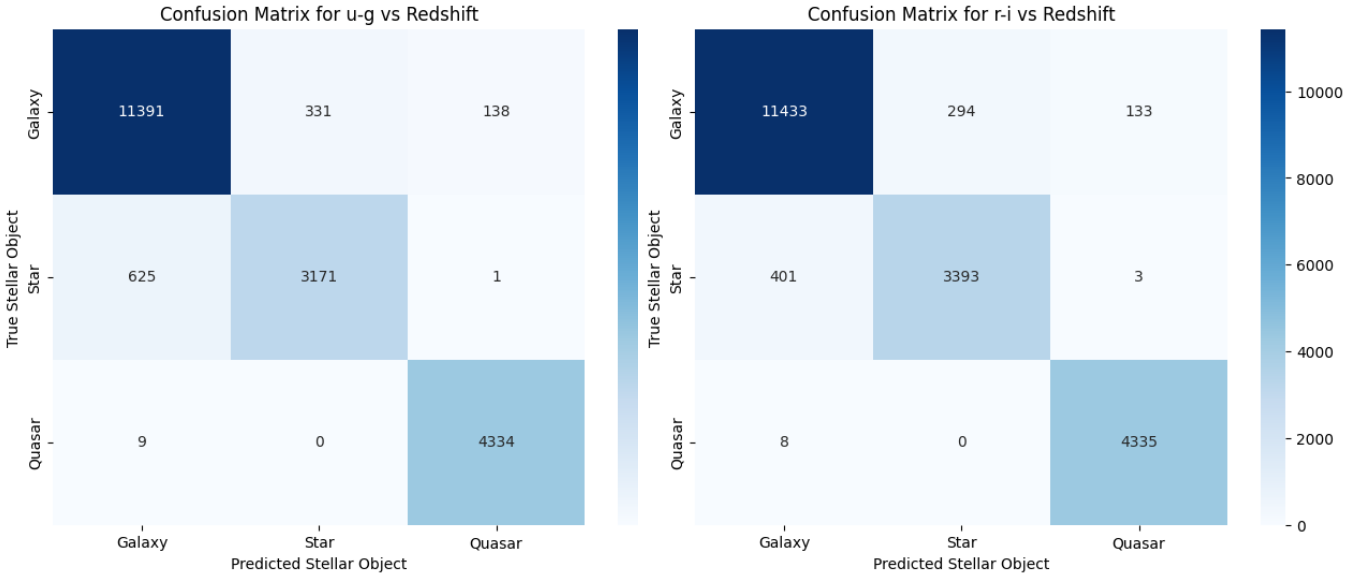
```
Accuracy using u-g against redshift: 0.944800

Accuracy using r-i against redshift: 0.956205
```

Figure 6: Manual Testing

3.3.2. Confusion Matrices:

The confusion matrices for both classification tasks are presented below as Fig a and Fig b. They visually represent the model's performance, displaying the number of correct and incorrect predictions for each class.



(a) u-g vs redshift classification

(b) r-i vs redshift classification

Figure 7: Confusion matrices

3.3.3. Manual Testing by User:

In my code I have also put a section to test the model manually. Whilst the extensive results do provide enough insight of the capabilities of the model, I felt a need to actually test the model myself.
The following is a snippet of the testing of the model manually.

```
kindly enter the u, g, i, r, and redshift values seperated by comma's: 22, 20,
19, 18, 17, 0.4
The predicted class is: GALAXY
kindly enter the u, g, i, r, and redshift values seperated by comma's: 21, 21.
2, 23, 14, 16.8, 0.46
The predicted class is: STAR
```

Figure 8: Manual Testing

In this case the predicted classes were correct. Given the high accuracy of the model (95%) It is very hard to get the model to predict falsely.

4. Conclusion

4.1. Achievements from the Classification

In this project, we embarked on an exploratory journey through the cosmos, employing the k-Nearest Neighbors (kNN) algorithm to classify celestial bodies into stars, galaxies, and quasars. As someone with a deep interest in astronomy, the opportunity to blend my passion with the practical application of machine learning has been very rewarding.
Through this classification effort, we have not only demonstrated the power of machine learning in interpreting complex astronomical data but also enhanced our understanding of the universe's vast expanse.
The accuracy achieved in distinguishing between different celestial objects using color indices and Redshift as features underscores the effectiveness of the kNN algorithm in handling such a nuanced task. By visualizing the results through confusion matrices, we've gained insightful perspectives on the model's performance, further fueling my fascination with the celestial realm.

Should the reader feel the need to test the model or refer to it, I have attached the link to the GitHub repository of my project:

Github repository: link to the project on Github

4.2. Learning Journey and Gratitude

My journey into the realms of data science and machine learning, particularly in applying these disciplines to astronomy, has been largely self-directed. Despite these topics not being covered in my college curriculum (yet), my curiosity led me to seek the knowledge. It was through engaging discussions with my seniors, who generously shared their expertise and experiences, that I was able to grasp the intricate concepts of the kNN algorithm and its application in classifying astronomical objects.

I am immensely grateful to my seniors for their guidance and support - especially for learning machine learning techniques. Furthermore, I would like to extend my heartfelt thanks to Professor Preeti. By granting us the freedom to choose our topics for the model prediction project, she not only fostered an environment of creativity and exploration but also allowed me to delve into the fascinating intersection of astronomy and machine learning.

As a student of Data Science and AI, I am excited about the prospects of further integrating machine learning techniques with the natural sciences, venturing deeper into this fascinating intersection of disciplines like astronomy.

Thank you

***