
Project Report: Classification of Stars, Galaxies, and Quasars (SDSS Data)

1. Introduction

Modern sky surveys such as the Sloan Digital Sky Survey (SDSS) collect large volumes of data containing photometric magnitudes (**u**, **g**, **r**, **i**, **z**), redshift, and coordinates for each observed source. Accurately classifying these objects into stars, galaxies, or quasars is crucial for astronomical research, enabling automated processing of massive catalogs.

In this project, I compare several machine learning methods and investigate how different features—particularly cosmological redshift—help distinguish objects at various distances. My specific goals are:

1. **Develop** and evaluate multiple classification models (Random Forest, CatBoost, SVM, MLP).
2. **Analyze** feature importance and the overall data structure via correlation heatmaps and clustering.
3. **Test** the hypothesis that redshift alone can effectively separate quasars from stars and galaxies.

---

2. Data and Preprocessing

I used a CSV file called star_classification.csv from SDSS, which includes:

- **Magnitudes** (u, g, r, i, z): Photometric measurements in different wavebands.
- **Redshift**: Indicates how far the spectrum is shifted due to the expansion of the universe (roughly linked to distance).
- **Coordinates** (alpha, delta): Right ascension and declination of each object.
- **Class** (class): The target label (Star, Galaxy, Quasar).

2.1. Data Cleaning and Preparation

- I removed technical or ID columns that do not carry useful information for classification.
- I encoded the class column using a LabelEncoder, mapping each class (Star, Galaxy, QSO) to a numeric label (e.g., 0, 1, 2).
- I split the data into training (80%) and test (20%) sets.
- All numeric features were standardized using StandardScaler to normalize magnitudes and redshift before training.

---

3. Methodology

3.1. Machine Learning Methods

1. **Random Forest**: An ensemble of decision trees. I used RandomizedSearchCV to optimize parameters (e.g., number of trees, max depth).

2. **CatBoost**: Gradient boosting on decision trees, performing well with minimal data preprocessing.
3. **SVM** (linear kernel): A classic support vector machine approach, with a linear kernel for simplicity.
4. **MLP**: A feedforward neural network (Multi-Layer Perceptron) with one hidden layer of 100 neurons.

3.2. Metrics and Evaluation

- **Accuracy** (proportion of correct predictions) on the test set.
- **Classification Reports**: Precision, recall, F1-score for each class (star, galaxy, quasar).
- **Confusion Matrices**: Reveal how objects are misclassified (which classes get mixed up).

3.3. Feature and Data Structure Analysis

- **Correlation Matrix**: To see how features are linearly related.
- **Feature Importance** (using RandomForest) to identify the most influential predictors.
- **K-Means Clustering**: To check the unsupervised structure of the data. I used the Elbow Method and Silhouette Score to estimate an optimal number of clusters and compared them with the original classes.

---

4. Results

4.1. Model Performance

After tuning hyperparameters via RandomizedSearchCV, **Random Forest** achieved high accuracy on the test set. **CatBoost** also performed strongly, often equaling or slightly surpassing the optimized RF model. Meanwhile, **SVM** and **MLP** reached robust accuracy but showed a bit more confusion between galaxies and quasars.

Typical test-set accuracies were approximately:

- **Random Forest**: ~95%
- **CatBoost**: ~94–95%
- **SVM**: ~93%
- **MLP**: ~92–94%

4.2. Confusion Matrices

- **Stars** are rarely misclassified (zero or near-zero redshift makes them easy to identify).
- **Galaxies vs. Quasars** can be more ambiguous. Quasars often have higher redshifts, but some lower-redshift quasars or high-redshift galaxies can overlap.
- Overall, the main source of classification error occurs between Galaxy and QSO.

4.3. Feature Importance

By examining the feature importances from RandomForest, I observed that **redshift** and certain photometric bands (e.g., r, i) contributed strongly to the classification. Meanwhile, **alpha** and **delta** (sky coordinates) had low importance — this is expected, since the position on the sky does not directly imply an object type.

4.4. Role of Redshift

A separate test using only the single feature redshift yielded accuracy around 80–85%, underscoring its strong predictive power:

- **Stars** (near zero)
- **Galaxies** (moderate redshift)
- **Quasars** (very high redshift, often above 2 or 3)

Thus, redshift alone effectively separates a large fraction of objects. Combining redshift with the photometric magnitudes (u, g, r, i, z) further improves the separation.

---

5. Discussion

1. **High Accuracy**: Multiple methods (especially Random Forest and CatBoost) achieve ~95% accuracy, suggesting that these features are highly informative.
2. **Galaxy–QSO Overlap**: While stars are almost always categorized correctly (redshift ~0), the boundary between galaxies and quasars can be blurred if their redshifts lie in overlapping ranges.
3. **Correlation**: Photometric bands correlate strongly with each other, yet redshift provides an additional dimension that helps differentiate classes.
4. **Coordinates**: The near-zero correlation of alpha/delta with the rest is expected in astronomy, as object type does not depend on position in the sky.

---

6. Conclusion

My investigation confirms that:

- **Redshift** is a key feature for distinguishing stars, galaxies, and quasars.
- **Ensemble methods** such as Random Forest and CatBoost produce consistently high performance.
- **SVM and MLP** also perform well but tend to produce slightly more galaxy–quasar confusion.
- **K-Means Clustering** reveals roughly 2–3 major clusters, aligning with the three classes but showing overlap between galaxies and quasars.
- **Coordinates** (alpha, delta) have minimal impact, which is logical in astronomy, since sky position alone does not determine object type.

---

7. Key Findings

1. **High Classification Accuracy** (~95%) across Star/Galaxy/QSO.

2. **Minimal Confusion for Stars**: Stars are almost never confused with the other classes, thanks to near-zero redshift.
3. **Main Errors** occur between Galaxy and Quasar, consistent with real astrophysical overlap for certain redshift values.
4. **Single-Feature Test**: Using only redshift yields around 80–85% accuracy.
5. **Coordinates** (alpha, delta) offer negligible classification power for object type.

In summary, this project demonstrates the effectiveness of modern machine learning approaches for SDSS object classification and emphasizes the critical importance of redshift for distinguishing galaxies from quasars.
