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**:
    - The photometric magnitudes (u, g, r, i, z) exhibit strong positive pairwise correlations. For instance, r and i can correlate at around 0.96, indicating that objects bright in r are usually bright in i.
    - Redshift shows moderate positive correlation with r (~0.43) and i (~0.49).
    - Alpha and delta (sky coordinates) have very low correlation with the other features, which is expected, since an object's position does not dictate its brightness or spectral shift.

Why might redshift correlate with r/i?

4. **Spectral Energy Distribution (SED) shift**: As objects are found at higher redshifts, certain rest-frame UV or blue features move into the redder filters (r, i).

5. **K-Corrections**: Bright emission lines can become more pronounced in the r/i wavebands when redshifted.

6. **Survey selection bias**: High-redshift objects that remain bright in r/i are more likely to be detected.

7. **Strong emission lines**: Some quasars or distant galaxies have lines that appear in the r/i bands once they are redshifted, boosting measured fluxes there.

- **Feature Importance**: Assessed primarily with RandomForest's feature_importances_.
- **K-Means Clustering**: Performed to check unsupervised structure (Elbow Method, Silhouette Score) and compare the number of discovered clusters to the original classes.

---

4. Results

4.1. Model Performance

After hyperparameter optimization (via RandomizedSearchCV for Random Forest), the final accuracies on the test set were:

- **Random Forest**: 0.97805

- **CatBoost**: 0.97775
- **MLP**: 0.9726
- **SVM**: 0.9593

Both Random Forest and CatBoost achieved exceptionally high performance, around **97.7–97.8%** accuracy. The MLP also performed well at **~97.26%**, while SVM followed at **~95.93%**.

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. However, given the high accuracy, these errors are relatively small.

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 — logical in astronomy, since 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%, highlighting 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 separation.

---

5. Discussion

1. **Exceptional Accuracy**: Random Forest and CatBoost both exceed 97% on the test set, indicating that the chosen features and models are highly effective.
2. **Galaxy–QSO Overlap**: While stars are almost always categorized correctly, the boundary between galaxies and quasars can be blurred if their redshifts lie in overlapping intervals.
3. **Correlation Insights**:
    - The **high correlation** among the filters (r vs. i) reflects astrophysical reality: bright objects in one optical band often appear bright in adjacent bands.
    - The **moderate positive correlation of redshift with r/i** suggests that high-redshift objects tend to remain bright in those filters, either due to spectral features shifting into r/i or due to survey selection that captures such objects.

4. **Coordinates**: Alpha and delta show negligible correlation with brightness, redshift, or class, which aligns with the fact that an object's location on the sky does not dictate its intrinsic properties.

---

6. Conclusion

My investigation confirms that:

- **Redshift** is a key feature for distinguishing stars, galaxies, and quasars.
- **Ensemble methods** (Random Forest, CatBoost) produce consistently high performance, each ~97.7–97.8% accuracy.
- **MLP** achieves about 97.26%, while **SVM** is around 95.93%, showing slightly more confusion between galaxies and quasars.
- **K-Means Clustering** indicates ~2–3 major clusters, aligning with the three classes but showing overlap in the Galaxy–QSO region.
- **Coordinates** (alpha, delta) have minimal impact, which is expected in astronomy, since sky position alone does not determine object type.

---

7. Key Findings

1. **Very High Classification Accuracy** (~97–98%) across Star/Galaxy/QSO.
2. **Minimal Confusion for Stars**: They are almost never mistaken for other classes, thanks to near-zero redshift.
3. **Main Errors** occur between Galaxy and Quasar, consistent with real astrophysical overlap for certain redshift ranges.
4. **Single-Feature Test**: Using only redshift yields around 80–85% accuracy.
5. **Moderate Positive Correlation** of redshift with r/i: Many high-redshift objects remain bright in these bands, helping models distinguish them.
6. **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.

.