AYDIN ADNAN MENDERES UNIVERSITY

ENGINEERING FACULTY

COMPUTER SCIENCE ENGINEERING DEPARTMENT

Data mining final project - Celestial Object Classification with SDSS DR19 Dataset

CSE418 Data Mining, Fall 2025-2026

Student's Name Surname:

Gökay Sepet - 201805068

Cemre Polat - 211805054

Özcan Erdem Tosun - 231805003

Lecturer:

Asst. Prof.Denizhan Demirkol

ABSTRACT

This study aims to perform automated classification of celestial objects (Star, Galaxy, and Quasar) using the Sloan Digital Sky Survey (SDSS) Data Release 19 (DR19). Within the scope of the study, a high quality dataset containing both photometric and spectroscopic data was constructed via custom SQL queries from the SDSS SkyServer . To enhance model performance, feature engineering techniques were applied to generate derived attributes such as color indices, Cartesian coordinates, and redshift derivatives in addition to raw features. To optimize dimensionality and maximize classification success, the most discriminative 12 features were selected using the ANOVA F test based SelectKBest method. Five supervised learning algorithms Random Forest, Gradient Boosting, Support Vector Machines (SVM), K Nearest Neighbors (KNN), and Naive Bayes were trained and evaluated using Accuracy, Precision, Recall, and F1-Score metrics. Experimental results demonstrated that the Gradient Boosting algorithm achieved the highest performance with an accuracy of 98.20%, followed by Random Forest with 98.15%. Furthermore, error analysis revealed that redshift information plays a critical role, particularly in distinguishing Quasars from other celestial objects.

1. INTRODUCTION

The primary objective of this project is to develop an automated classification system capable of accurately distinguishing between three fundamental types of celestial objects: Stars, Galaxies, and Quasars (QSO). To achieve this, the study analyzes a rich dataset derived from the Sloan Digital Sky Survey (SDSS) Data Release 19 (DR19), which provides comprehensive photometric and spectroscopic observations of the universe.

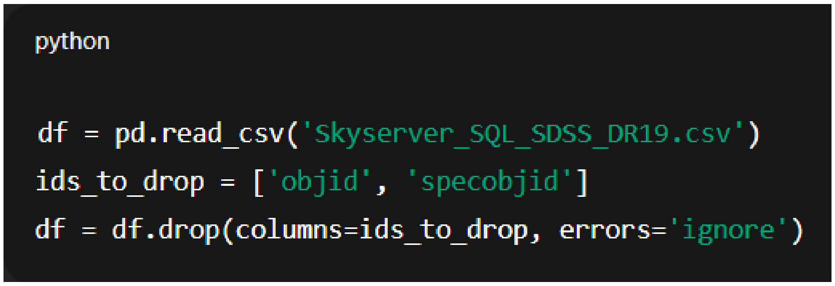
With the exponential growth of astronomical data in modern sky surveys, manual classification by astronomers has become impractical, time-consuming, and prone to human error. Therefore, the application of machine learning algorithms is essential to process these vast datasets efficiently and with high precision. This study specifically aims to demonstrate that the integration of domain-specific **feature engineering** combined with **ANOVA-based feature selection** significantly enhances the classification accuracy and robustness of machine learning models compared to using raw data alone. This study follows the **KDD (Knowledge Discovery in Databases)** process, encompassing data selection, preprocessing, transformation, data mining (using 5 different algorithms), and evaluation. The study follows the standard data mining pipeline:

- data acquisition from an external scientific database
- data cleaning and feature engineering
- feature selection
- model training with multiple algorithms
- evaluation and visual error analysis

All experiments were implemented in Python using pandas, numpy, matplotlib, seaborn, and scikit-learn.

1.1 DATASET AND TARGET VARIABLE

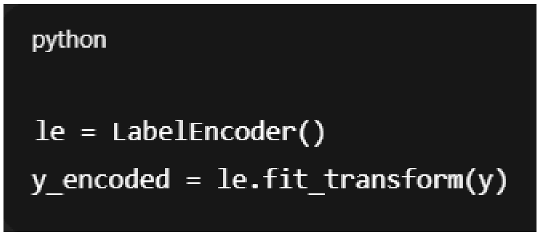
The working dataset is stored locally as Skyserver_SQL_SDSS_DR19.csv.
After loading the CSV file:



```python
df = pd.read_csv('Skyserver_SQL_SDSS_DR19.csv')
ids_to_drop = ['objid', 'specobjid']
df = df.drop(columns=ids_to_drop, errors='ignore')
```

- **Independent variables (X):** all columns except class.
- **Target variable (y):** the class column, containing the physical type of each object.

Since class is categorical, it is encoded to integers using LabelEncoder:



```python
le = LabelEncoder()
y_encoded = le.fit_transform(y)
```

The encoded labels correspond to the three main object classes (STAR, GALAXY, QSO).

2. LITERATURE REVIEW

The classification of celestial objects has undergone a paradigm shift from manual inspection of photographic plates to automated processing of massive digital datasets. The Sloan Digital Sky Survey (SDSS) has been a cornerstone in this transition, providing one of the most comprehensive 3D maps of the universe.

2.1. The SDSS Photometric System

The foundational technical framework of the SDSS was established by ***York et al. (2000)***, who described the survey's telescope design and imaging strategy to map a quarter of the sky. A critical component of this survey is its photometric system, which was defined by ***Fukugita et al. (1996)***. They introduced the five-band filter system *(u, g, r, i, z)* that covers the optical spectrum from ultraviolet to infrared, enabling precise color measurements of celestial bodies.

Over the years, the SDSS has released incremental data updates. While earlier studies utilized data from the Seventh Data Release (DR7) as described by ***Abazajian et al. (2009),*** recent research has shifted towards more advanced releases. For instance, ***(Ahumada et al., 2020)*** detailed the improvements in the 16th Data Release (DR16), including better calibration and expanded spectral libraries. This study utilizes the Data Release 19 (DR19), which represents the most current and refined iteration of these datasets.

2.2. Machine Learning Approaches in Astronomy

With the exponential growth of data, traditional statistical methods have been replaced by machine learning (ML) algorithms. Numerous studies have benchmarked various classifiers on SDSS data.

Decision Trees, Random Forests, and Gradient Boosting: Ensemble methods have proven highly effective in astronomical classification. *(Arafat et al., 2025; Solorio-Ramírez et al., 2023)* demonstrated that Random Forest classifiers are robust against noise and missing data, often achieving accuracy rates exceeding 95%. Similarly***, (Ball et al., 2006; Vasconcellos et al., 2011)*** used decision tree ensembles to separate stars from galaxies, highlighting the importance of morphological features. In addition to these, Gradient Boosting Machines (GBM), particularly implementations like XGBoost and LightGBM, have recently set new benchmarks in the field. These methods employ a boosting technique that builds trees sequentially to minimize residual errors, proving exceptionally powerful in handling imbalanced astronomical datasets and capturing complex non-linear relationships between celestial features.

- **Support Vector Machines (SVM) and KNN:** The efficacy of Support Vector Machines (SVM) in high-dimensional spaces has been explored by several researchers. ***(Peng et al., 2013; Suwaid et al., 2025)*** noted that while SVMs are computationally intensive, they provide superior decision boundaries for non-linear data distributions. On the other hand, K-Nearest Neighbors (KNN) serves as a strong baseline. ***(Ashai et al., 2022; Lin, 2023)*** applied KNN to photometric data and found that while simple, it is sensitive to the choice of 'k' and the distance metric used.

- **Comparison of Algorithms:** Comparative studies often highlight trade-offs between speed and accuracy. For example, (Buisson et al., 2015; Hickey et al., 2023; Vavilova et al., 2021) compared Naive Bayes, Neural Networks, and SVM, concluding that tree-based models generally offer the best balance of interpretability and performance for tabular astronomical data.

2.3. Feature Engineering and Selection

The success of classification models depends heavily on feature selection. (Donalek et al., 2013; Kamath et al., 2004; Zheng & Zhang, 2008) emphasized that using raw magnitudes alone is often insufficient. instead, "color indices" *(differences between magnitudes like u-g or g-r)* are standard inputs in modern pipelines as they correlate directly with stellar temperature and galaxy type.

Furthermore, (Fan, 2006; Greiner et al., 2021; Hewett, 2010; Warren & Hewett, 1990) highlighted that **redshift** is the single most discriminative feature for identifying Quasars (QSO), which are characterized by high recession velocities. Recent studies typically implement software libraries such as **Scikit-learn** ***(Pedregosa et al., 2011)*** and **Pandas** ***(McKinney, 2010)*** to facilitate these complex feature engineering and modeling tasks efficiently.

3. METHODOLOGY

The project structure is aligned with the CRISP-DM framework, starting from Business/Data Understanding to the final Evaluation of 5 distinct machine learning models.

3.1 Data Acquisition Methodology

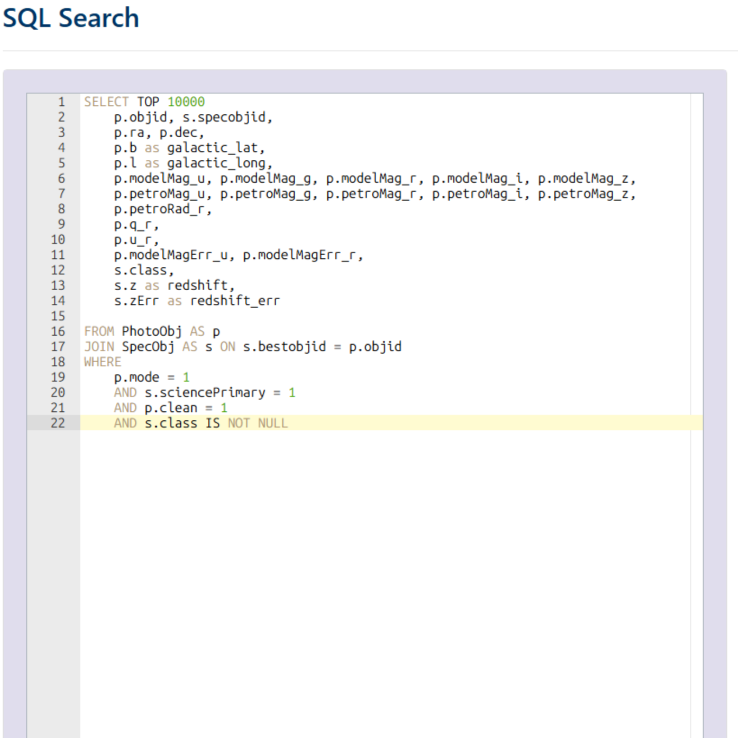
Unlike common pre-packaged datasets from platforms like Kaggle, this dataset was constructed directly from the **SDSS DR19** database.
The **SkyServer SQL Search** interface was used to run a custom SQL query that joins:

- PhotoObj (photometric objects): image-based quantities such as magnitudes, radii and shape parameters,
- SpecObj (spectroscopic objects): precise spectroscopic redshift and spectral classifications.

The query applied several quality filters:

- p.mode = 1 → selects primary detections,
- p.clean = 1 → ensures that only clean, well-processed photometric measurements are used,
- s.sciencePrimary = 1 → keeps the main, science-grade spectroscopic observation,
- s.class IS NOT NULL → excludes objects without a valid spectroscopic class.

This design guarantees that each row represents a **high-quality primary observation** with reliable photometry and spectroscopy.



```
SELECT TOP 10000
    p.objid, s.specobjid,
    p.ra, p.dec,
    p.b as galactic_lat,
    p.l as galactic_long,
    p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
    p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
    p.petroRad_r,
    p.q_r,
    p.u_r,
    p.modelMagErr_u, p.modelMagErr_r,
    s.class,
    s.z as redshift,
    s.zErr as redshift_err

FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
    p.mode = 1
    AND s.sciencePrimary = 1
    AND p.clean = 1
    AND s.class IS NOT NULL
```

Thanks to this custom query, the resulting table not only contains standard magnitudes and redshifts, but also richer morphological information such as Petrosian magnitudes, axis ratios and Stokes parameters. This provides a more expressive feature space for the classification problem.

3.2 Data Dictionary

The final dataset consists of:

1. **Raw features** obtained directly from the SQL query, and
2. **Engineered features** computed from these raw attributes to better capture physical structure.

3.2.1 Raw Features (Directly from SDSS)

These variables correspond to physical measurements recorded by the telescope:

- **ra, dec**
    - Right Ascension and Declination (J2000, degrees).
    - Sky coordinates analogous to longitude and latitude on Earth.
- **modelMag_[u, g, r, i, z]**
    - *Model magnitudes* in five photometric bands (Ultraviolet, Green, Red, Near-IR, IR).
    - Optimized for point sources and therefore very informative for **Stars**.
- **petroMag_[u, g, r, i, z]**
    - *Petrosian magnitudes* in the same bands.
    - Better capture the extended light profiles of **Galaxies**, including flux from their outer regions.
- **redshift (z)**
    - Spectroscopic redshift, measuring how much an object's light is stretched due to cosmic expansion.
    - A key discriminator for **Quasars (QSO)**, which typically have high redshift.
- **redshift_err**
    - Uncertainty associated with the redshift measurement.
- **petroRad_r**
    - Petrosian radius in the r-band, representing an angular size indicator.
- **q_r (axis ratio)**
    - Minor-to-major axis ratio in the r-band.
    - Values near 1.0 correspond to nearly round objects; smaller values indicate more elongated galaxies.

- **u_r (Stokes parameter)**
    - Shape-related parameter describing distortion and alignment.
- **galactic_lat (b) and galactic_long (l)**
    - Galactic latitude and longitude, i.e., positions with respect to the Milky Way plane and center.
- **class (target)**
    - Spectroscopic class: **STAR, GALAXY, QSO**.

3.2.2 Engineered Features

To improve separability and incorporate domain knowledge, several derived features were added:

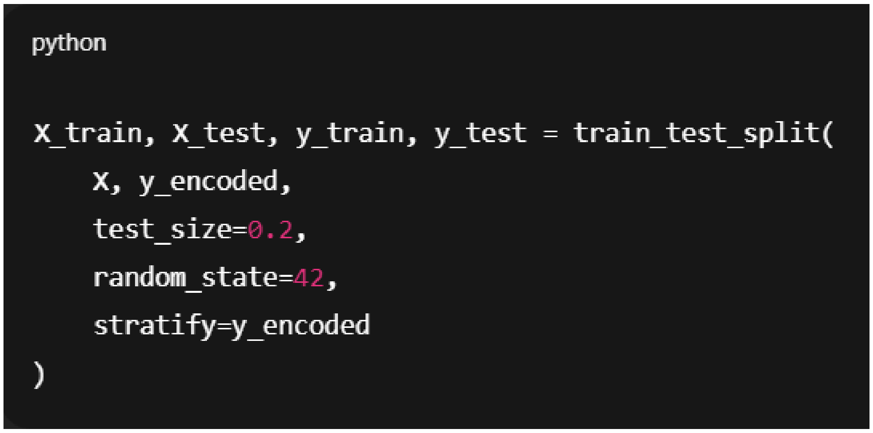
- **Color indices (Model magnitudes)**
    - u_g_model, g_r_model, r_i_model, i_z_model
    - Differences between adjacent bands using modelMag values.
    - Capture the spectral shape and are physically related to temperature and composition.
- **Color indices (Petrosian magnitudes)**
    - u_g_petro, g_r_petro, r_i_petro, i_z_petro
    - Same idea as above but computed from petroMag.
    - Especially useful for extended galaxies whose outskirts are better captured by Petrosian flux.
- **Cartesian coordinates**
    - cartesian_x, cartesian_y, cartesian_z
    - Conversion of spherical (ra, dec) coordinates into 3D Cartesian space for geometric analysis.
- **Redshift derivatives**
    - redshift_sq = z^2 – amplifies separation at higher redshift values.
    - redshift_snr – redshift weighted by its reliability (signal-to-noise), so high-error measurements have lower effective contribution.
- **Morphological interaction term**
    - q_r_sq = $(\text{axis ratio})^2$ – emphasizes differences between perfectly round sources (typical for stars) and flattened disk galaxies.

These engineered features are later used by the **SelectKBest** procedure to identify the most discriminative subset of variables for the models.

3.3 Methodology: Pre-processing and Modelling

3.3.1 Train/Test Split (Hold-Out)

To evaluate generalization performance, the data is split into training and test sets with an 80/20 ratio, preserving class balance:

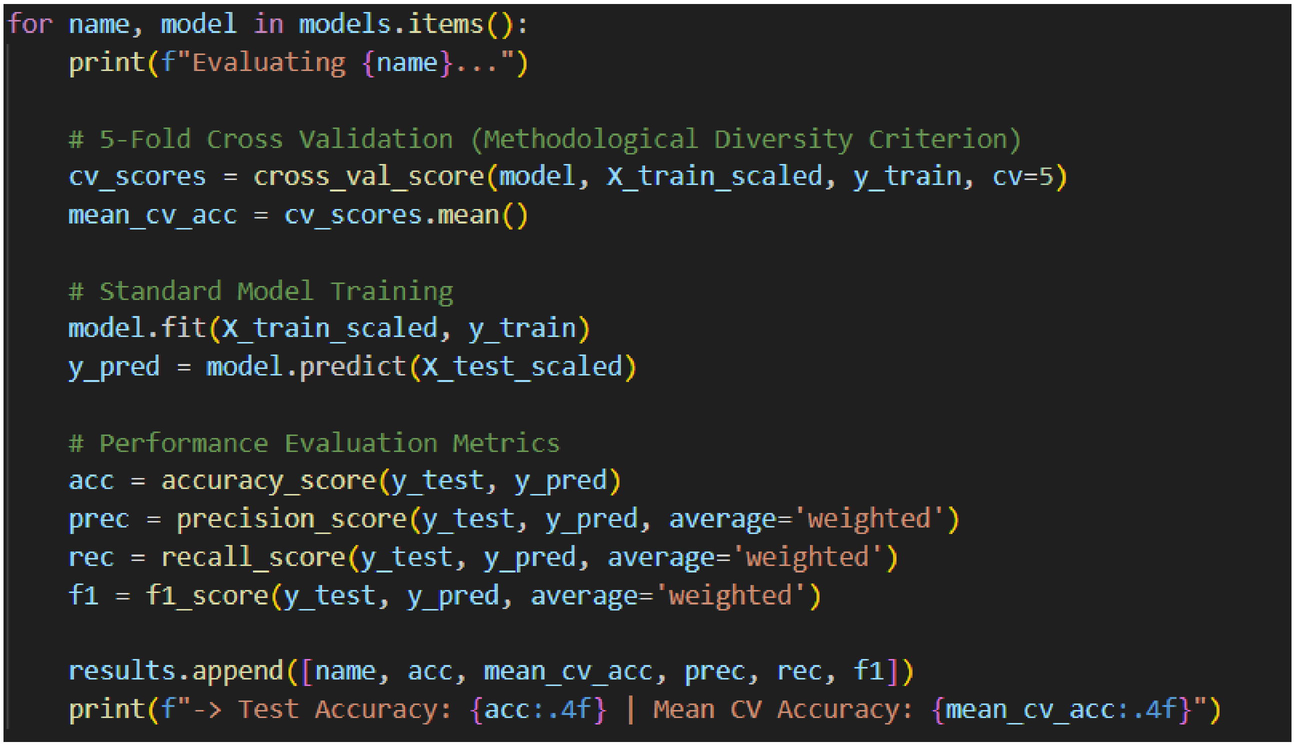


python

```python
X_train, X_test, y_train, y_test = train_test_split(
    X, y_encoded,
    test_size=0.2,
    random_state=42,
    stratify=y_encoded
)
```

3.3.2 Cross Validation (CV)

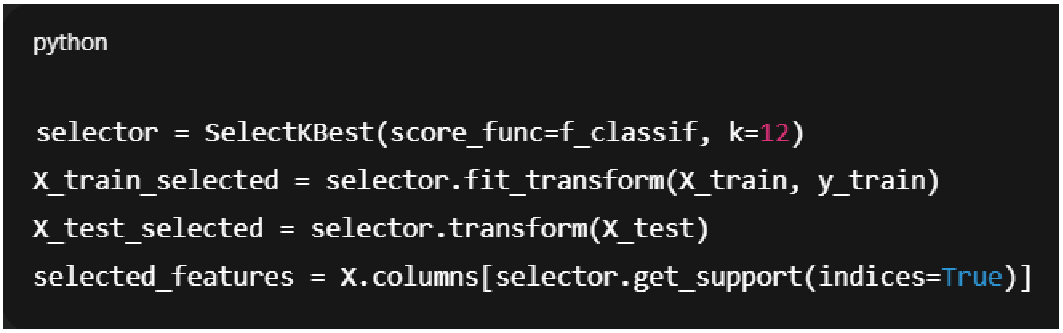
To ensure model robustness and prevent overfitting, 5-Fold Cross-Validation was employed alongside the 80/20 hold-out split:



```python
for name, model in models.items():
    print(f"Evaluating {name}...")

    # 5-Fold Cross Validation (Methodological Diversity Criterion)
    cv_scores = cross_val_score(model, X_train_scaled, y_train, cv=5)
    mean_cv_acc = cv_scores.mean()

    # Standard Model Training
    model.fit(X_train_scaled, y_train)
    y_pred = model.predict(X_test_scaled)

    # Performance Evaluation Metrics
    acc = accuracy_score(y_test, y_pred)
    prec = precision_score(y_test, y_pred, average='weighted')
    rec = recall_score(y_test, y_pred, average='weighted')
    f1 = f1_score(y_test, y_pred, average='weighted')

    results.append([name, acc, mean_cv_acc, prec, rec, f1])
    print(f"-> Test Accuracy: {acc:.4f} | Mean CV Accuracy: {mean_cv_acc:.4f}")
```

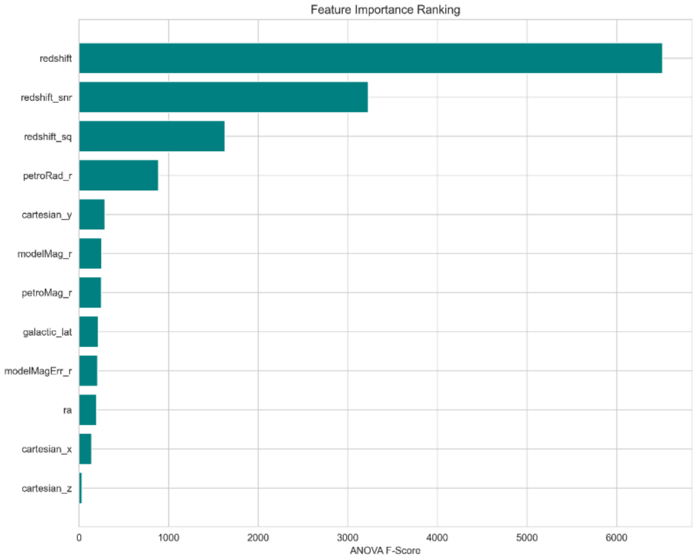
3.3.3 Feature Selection with ANOVA F-test

Instead of using all features, the project applies **SelectKBest** with ANOVA F-test to keep only the 12 most informative attributes:



```python
selector = SelectKBest(score_func=f_classif, k=12)
X_train_selected = selector.fit_transform(X_train, y_train)
X_test_selected = selector.transform(X_test)
selected_features = X.columns[selector.get_support(indices=True)]
```

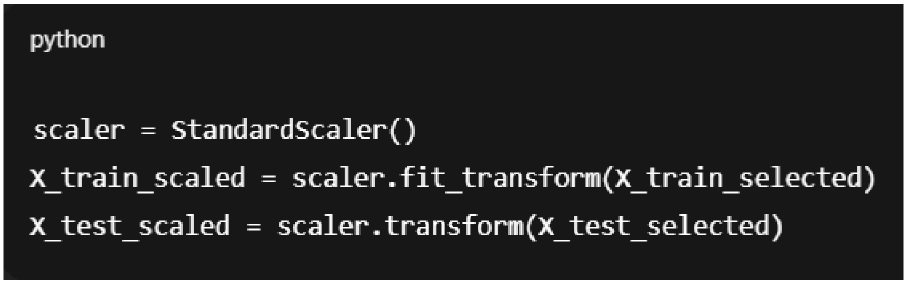
This step reduces dimensionality, removes noisy variables and focuses the models on the most discriminative aspects of the data.



The bar chart shows that color indices and redshift-related features rank very high, confirming their physical relevance for separating QSOs from stars and galaxies.

3.3.4 Feature Scaling

Since several algorithms are sensitive to the scale of input features, a **StandardScaler** is applied on the selected features:

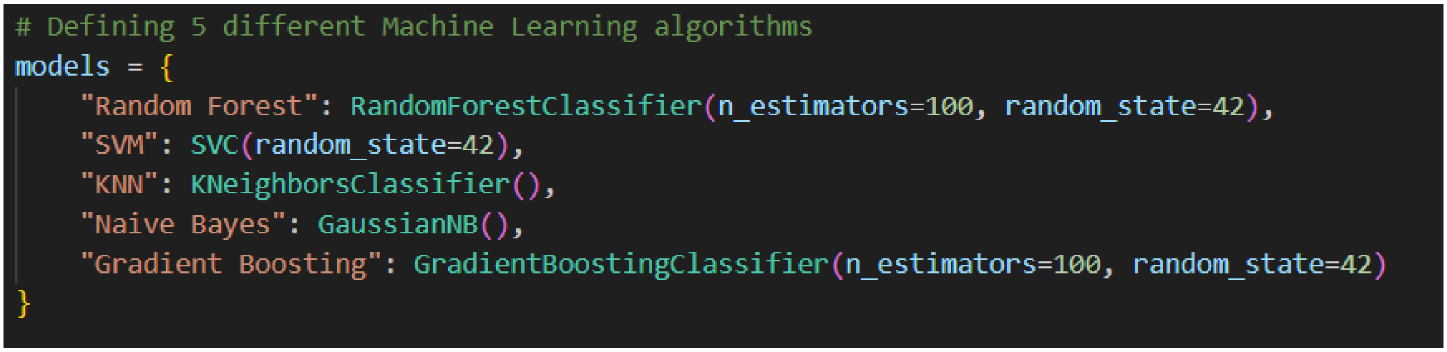


```python
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_selected)
X_test_scaled = scaler.transform(X_test_selected)
```

This transformation centers each feature to zero mean and unit variance.

3.3.4 Classification Algorithms

Five supervised learning algorithms are trained on the scaled, selected features:



```
# Defining 5 different Machine Learning algorithms
models = {
    "Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
    "SVM": SVC(random_state=42),
    "KNN": KNeighborsClassifier(),
    "Naive Bayes": GaussianNB(),
    "Gradient Boosting": GradientBoostingClassifier(n_estimators=100, random_state=42)
}
```

For each model, the following metrics are computed on the test set:

- Accuracy
- Precision (weighted)
- Recall (weighted)
- F1-Score (weighted)

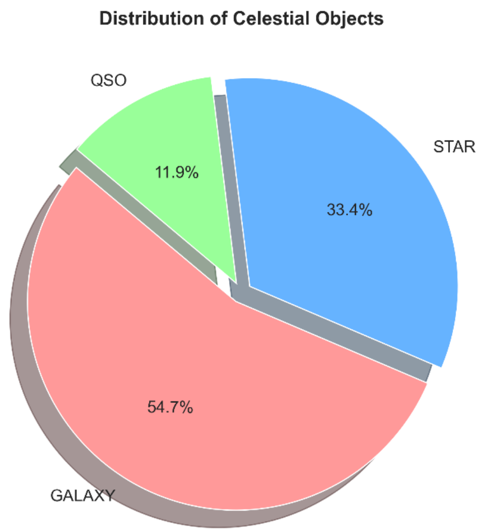
---

4. RESULTS

The second part of the script focuses on exploratory plots to better understand the data distribution and class separability.

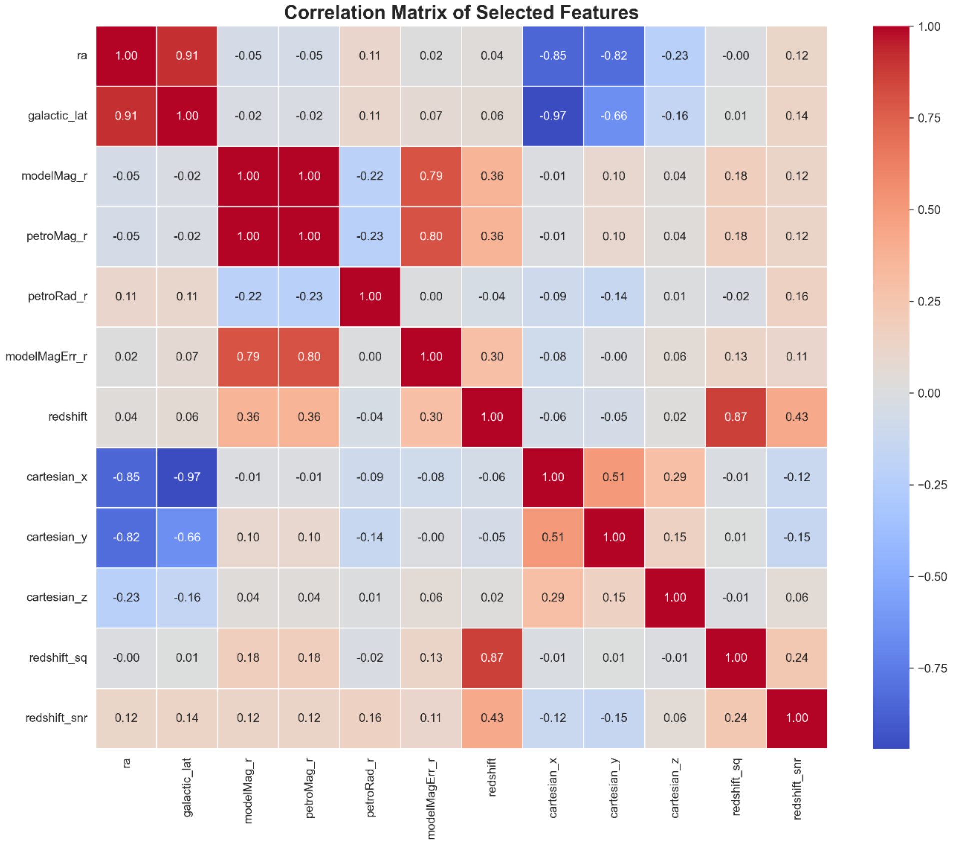
4.1 Exploratory Data Analysis (EDA)

4.1.1 Class Distribution (Pie Chart)



The pie chart shows that **GALAXY** and **STAR** objects constitute the majority of the dataset, while **QSO** makes up a noticeably smaller portion.
This imbalance indicates a potential **imbalanced dataset problem**.
Therefore, model performance is not evaluated solely with overall accuracy; **F1-Score** and other weighted metrics are also considered to avoid being biased towards the majority classes.

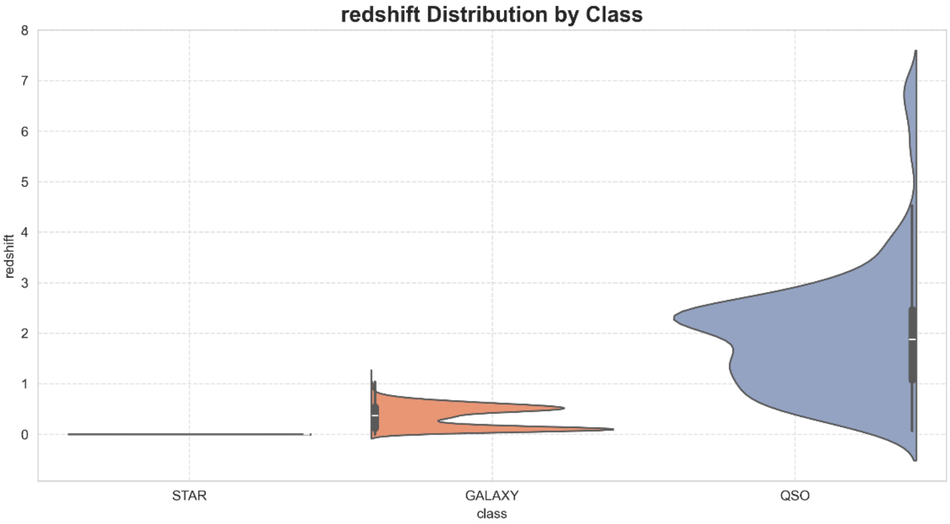
4.1.2 Correlation Heatmap of Selected Features



The heatmap reveals strong positive correlations among several magnitude-related features. This confirms that different brightness measurements in similar bands carry overlapping information.

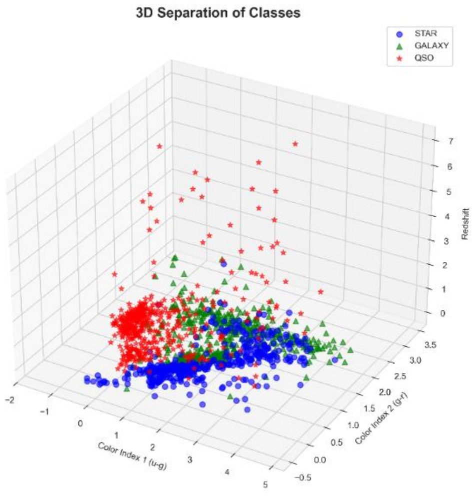
For this reason, the **feature engineering step focuses on color indices (e.g. u–g, g–r)**, which capture differences between bands and tend to be more discriminative than the raw magnitudes themselves.

4.1.3 Redshift Distribution by Class (Violin Plot)



The violin plot clearly shows that QSOs have a much wider and higher redshift distribution compared to stars and galaxies. While stars cluster around very small redshift values and galaxies occupy an intermediate range, QSOs extend to significantly larger redshifts. This visual evidence explains why redshift and its derived features appear as dominant variables in the feature selection ranking.

4.1.4 3D Scatter Plot: Color Indices vs. Redshift



In this plot, two color indices (e.g. u–g and g–r) and redshift are used as the three axes. Each class is visualized with different colors and markers, with up to 500 samples per class for clarity.

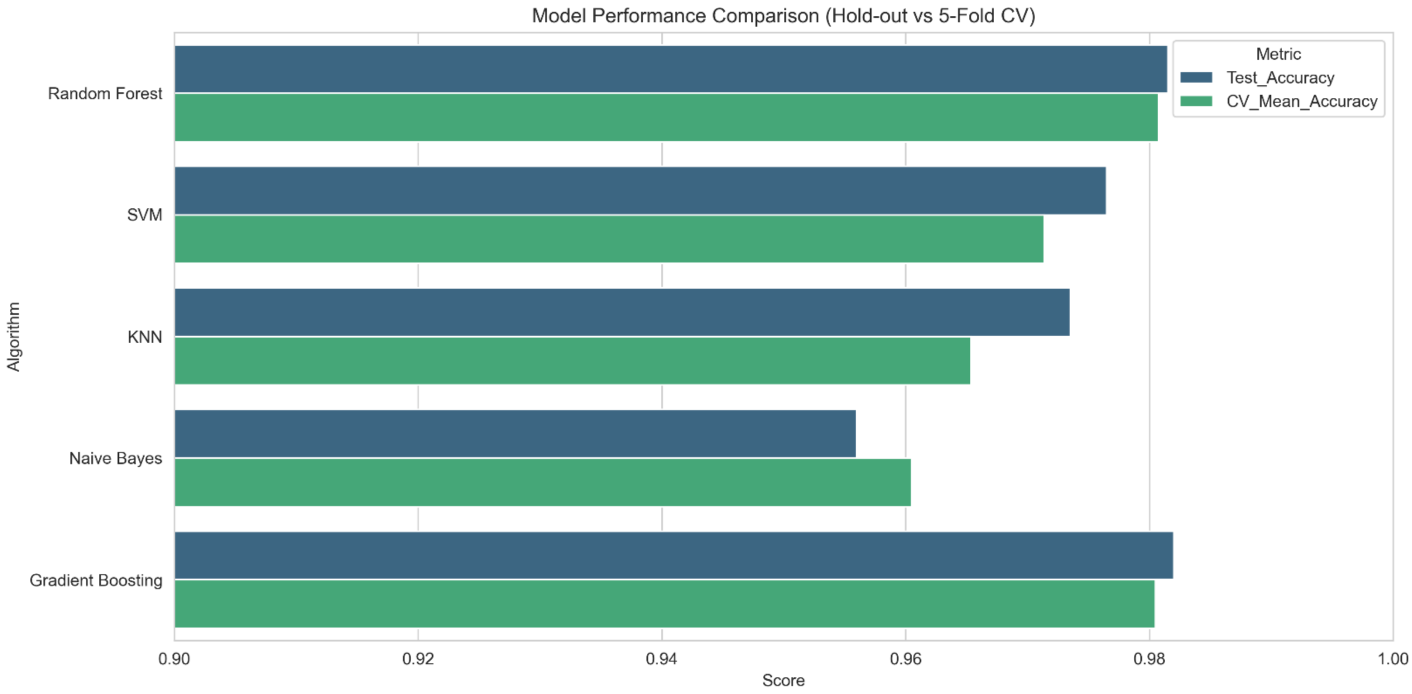
The figure shows that:

- QSOs (usually red points) tend to form a separate cloud at higher redshift levels,
- stars and galaxies overlap more in some regions of the color–redshift space.

This overlap visually explains why simpler distance-based methods such as KNN can occasionally misclassify borderline objects, while more complex models (e.g. Random Forest, SVM) benefit from a richer, non-linear decision boundary.

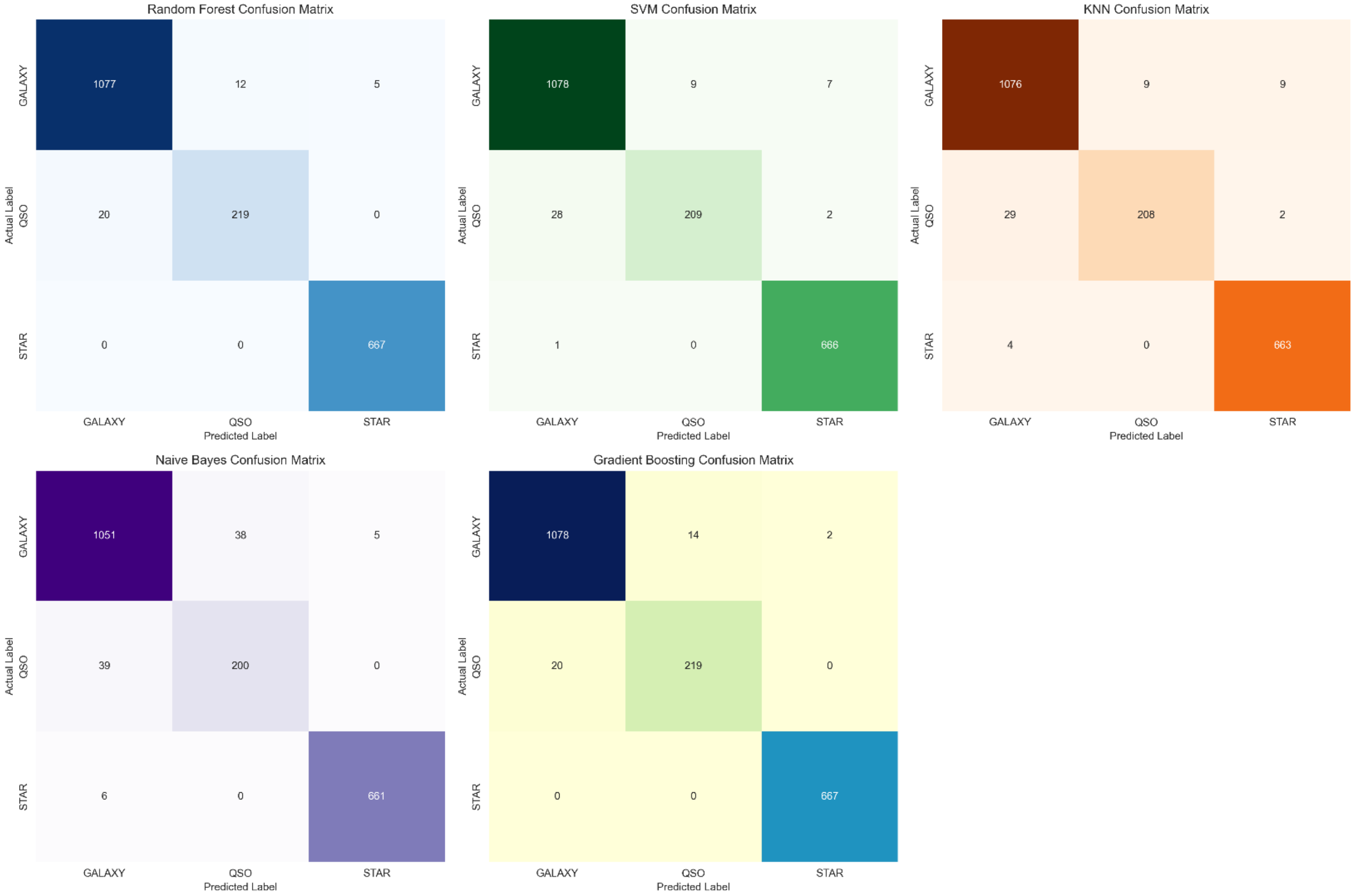
5. Experimental Results and Discussion

5.1 Accuracy Comparison of Algorithms



The bar chart summarises the overall accuracy of the four algorithms.
All methods achieve relatively high accuracy, but Gradient Boosting stand out with the best scores.

5.2 Confusion Matrices (Error Analysis)



This figure contains a 2×3 grid of confusion matrices:

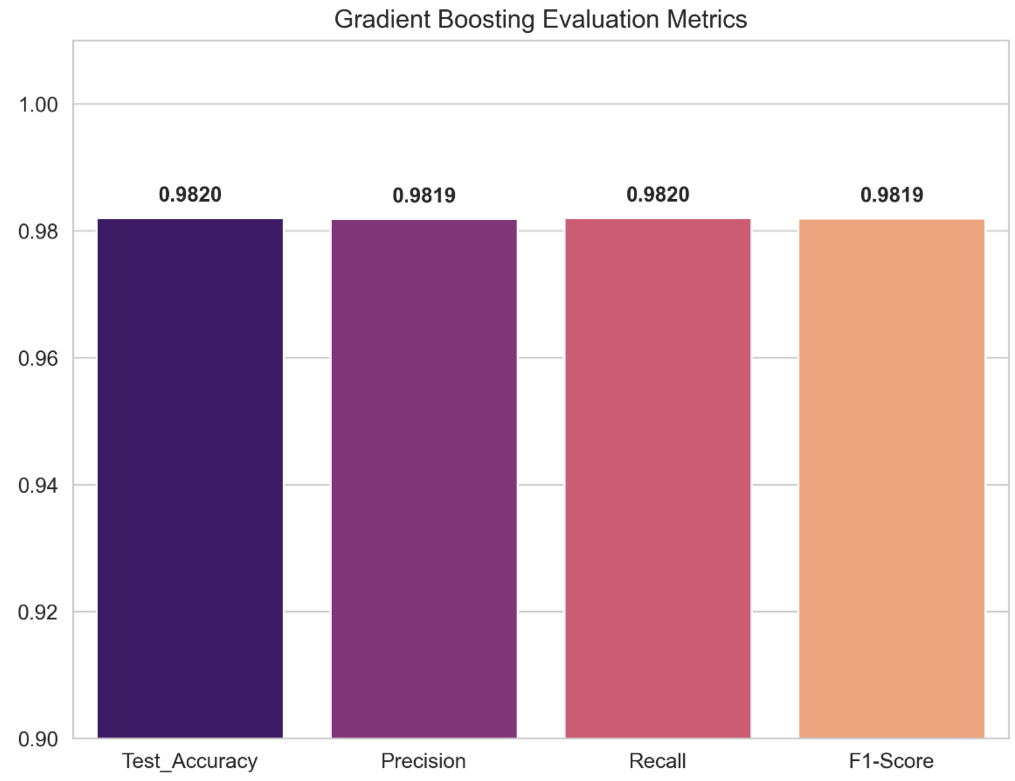
- top-left: Random Forest
- top-middle : SVM
- top-right: KNN
- bottom-left: Naive Bayes
- bottom-middle: Gradient Boosting

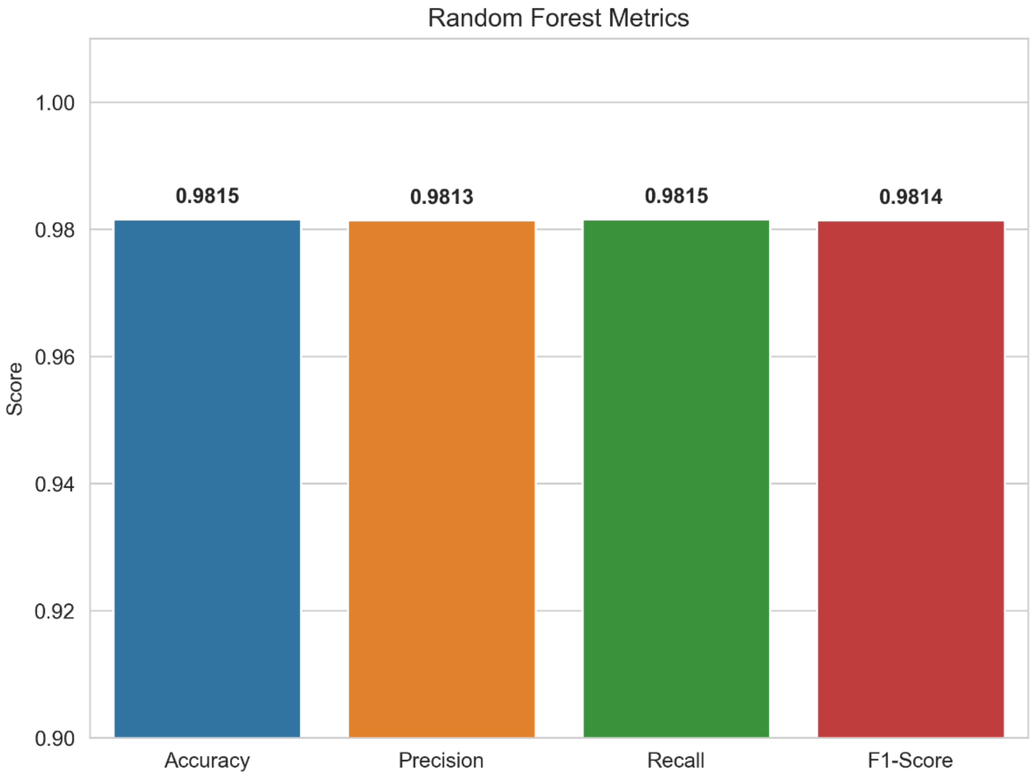
By inspecting these matrices, we can see:

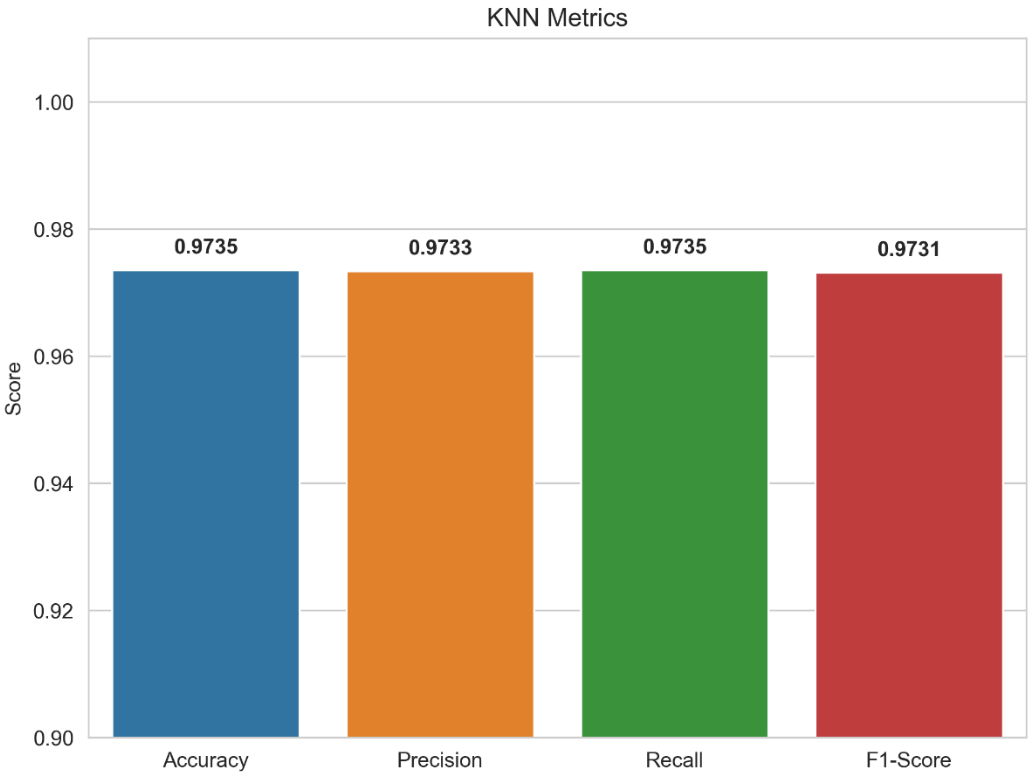
- which classes are most often confused,
- whether a model tends to over-predict a particular class,
- whether minority classes such as QSO are correctly identified.

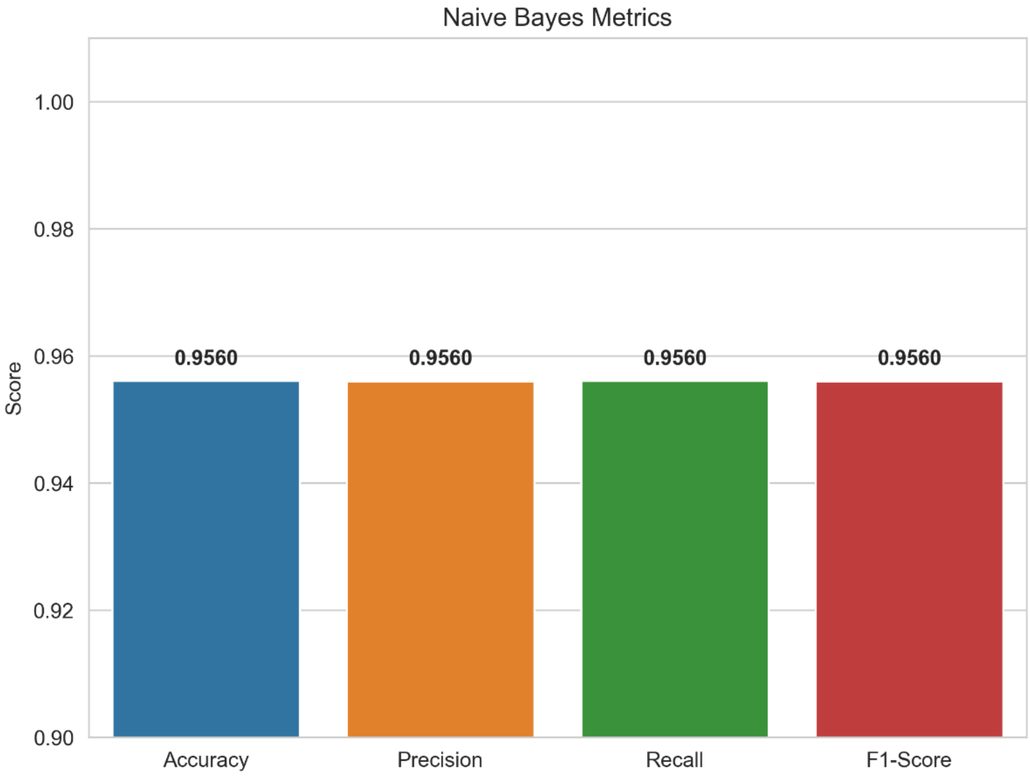
Typically, misclassifications occur between **GALAXY and QSO**, which is physically reasonable given that both can have extended structures and overlapping color properties at certain redshifts.

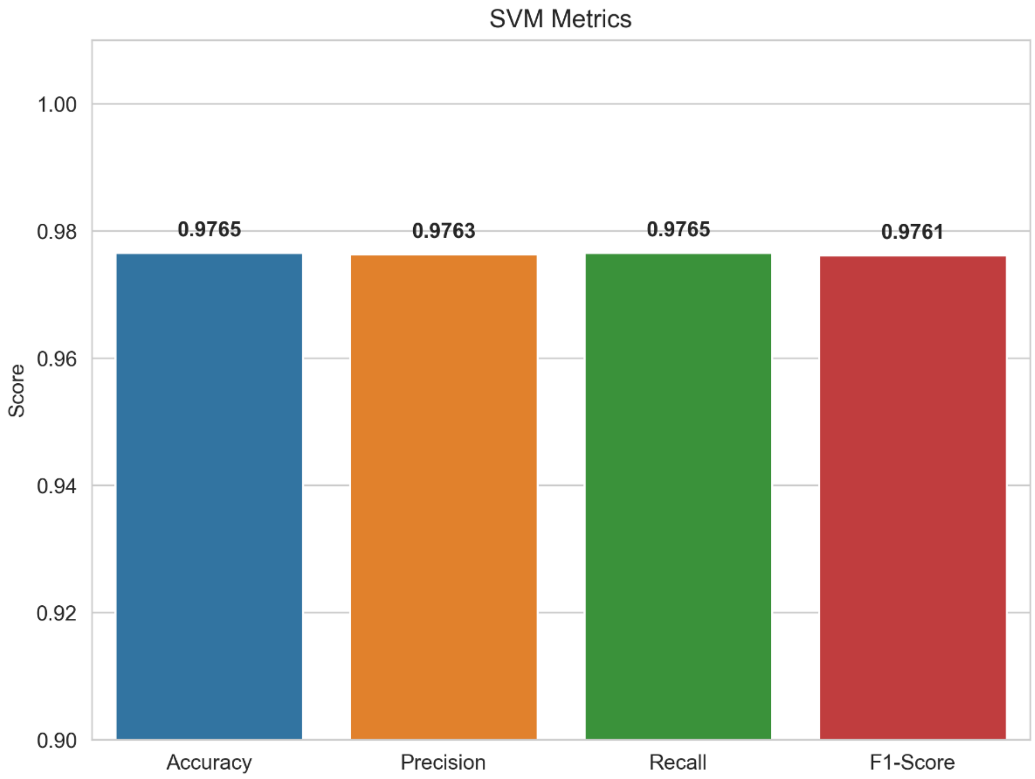
5.3 Per-Algorithm Metric Profiles






KNN Metrics
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.9735
0.9733
0.9735
0.9731
Accuracy
Precision
Recall
F1-Score


Naive Bayes Metrics
Score
1.00
0.98
0.96
0.94
0.92
0.90
0.9560
0.9560
0.9560
0.9560
Accuracy
Precision
Recall
F1-Score

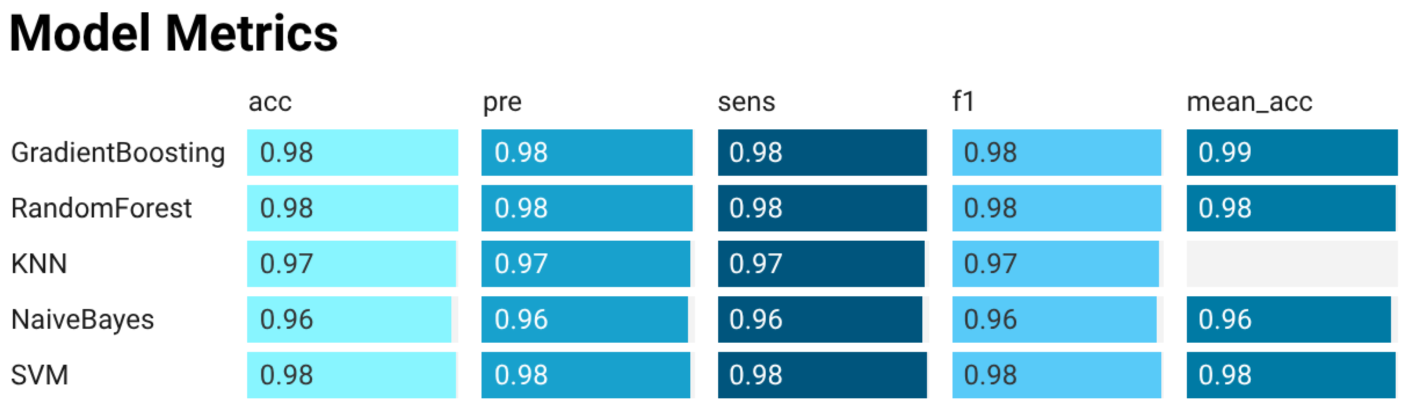


For each algorithm, accuracy, precision, recall and F1-Score are plotted side by side.
The y-axis is restricted to [0.90, 1.01], which makes small differences visible.
From these charts we can conclude:

- some models have slightly higher **precision** (fewer false positives),
- others achieve better **recall** on minority classes,
- the **F1-Score** balances these two aspects and is used as the main comparison metric in the presence of class imbalance.

In addition to tree-based and distance-based methods, **Gradient Boosting Classifier** was implemented to leverage boosting techniques for minimizing classification errors.

5.3 Overall Metrics



Model Metrics

	acc	pre	sens	f1	mean_acc
GradientBoosting	0.98	0.98	0.98	0.98	0.99
RandomForest	0.98	0.98	0.98	0.98	0.98
KNN	0.97	0.97	0.97	0.97	
NaiveBayes	0.96	0.96	0.96	0.96	0.96
SVM	0.98	0.98	0.98	0.98	0.98

6. Conclusion

In this project, an SDSS DR19-based dataset of celestial objects was constructed via a custom SQL query, enriched with engineered features, and used to solve a multi-class classification problem. The study demonstrated that integrating domain knowledge (astrophysical color indices and redshift derivatives) with automated feature selection significantly improves model performance. A **carefully curated dataset** of high-quality photometric and spectroscopic observations using SDSS SkyServer.

The main contributions of this work are:

- A set of **physically meaningful engineered features** (color indices, redshift derivatives, morphological interactions).
- A comparative study of five classical machine learning algorithms with **feature selection** via ANOVA F-test.
- A thorough **exploratory analysis** that visually explains where and why certain misclassifications occur.

6.1 Limitations

Despite the high accuracy achieved, this study has certain limitations. The dataset exhibited a significant class imbalance, with Quasars (QSO) being underrepresented compared to Stars and Galaxies . Although weighted evaluation metrics (Weighted Precision, Recall, and F1-Score) were utilized to mitigate this bias, collecting a larger volume of data specifically targeting minority classes could further validate the robustness of the model.

6.2 Future Work

For future research, Deep Learning architectures, such as Multi-Layer Perceptrons (MLP) or Convolutional Neural Networks (CNNs), could be explored to analyze spectral data more effectively. Additionally, incorporating more advanced attributes, such as temporal variability features or infrared data from other surveys, could provide new dimensions for classification and further minimize error rates.

6.3 ACKNOWLEDGES

We would like to thank our lecturer Asst. Prof. Denizhan Demirkol for his guidance throughout the course, and our teammates for their dedicated collaboration. The entire team contributed equally to the project.

7. References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Prieto, C. A., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A. L., Barentine, J. C., Bassett, B. A., Becker, A. C., Beers, T. C., Bell, E. F., Belokurov, V., Berlind, A. A., Berman, E. F., … Zucker, D. B. (2009). THE SEVENTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY. *The Astrophysical Journal Supplement Series*, *182*(2), 543–558. https://doi.org/10.1088/0067-0049/182/2/543

Acharya, V., Bora, P. S., Navin, K., Nazareth, A., Anusha, P. S., & Rao, S. (2018). Classification of SDSS photometric data using machine learning on a cloud. *Current Science*, *115*(2), 249–257.

Ahumada, R., Prieto, C. A., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., Aubert, M., Avila, S., Avila-Reese, V., Badenes, C., Balland, C., Barger, K., Barrera-Ballesteros, J. K., Basu, S., Bautista, J., Beaton, R. L., … Zou, H. (2020). The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *The Astrophysical Journal Supplement Series*, *249*(1), 3. https://doi.org/10.3847/1538-4365/ab929e

Arafat, Y., Begum, R., Rahman, M. S., & Kibria, M. K. (2025). Star Classification Using Machine Learning: A Comparative Analysis of Random Forest and LightGBM on SDSS Data. *International Journal of Statistical Sciences*, *25*(2), 159–172. https://doi.org/10.3329/ijss.v25i2.85778

Ashai, M., Mukherjee, R. G., Mundharikar, S. P., Kuanr, V. D., & Harikrishnan, R. (2022). Classification of Astronomical Objects using KNN Algorithm. In V. Bhateja, S. C. Satapathy, C. M. Travieso-Gonzalez, & T. Adilakshmi (Eds.), *Smart Intelligent Computing and Applications, Volume 1* (pp. 377–387). Springer Nature. https://doi.org/10.1007/978-981-16-9669-5_34

Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. (2006). Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. *The Astrophysical Journal*, *650*(1), 497. https://doi.org/10.1086/507440

Buisson, L. du, Sivanandam, N., Bassett, B. A., & Smith, M. (2015). *Machine learning classification of SDSS transient survey images | Monthly Notices of the Royal Astronomical Society | Oxford Academic*. https://academic.oup.com/mnras/article/454/2/2026/1051683

Donalek, C., Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Drake, A. J., Kumar, A. A., Philip, N. S., Fuchs, T. J., Turmon, M. J., Yang, M. T.-C., & Longo, G. (2013). Feature selection strategies for classifying high dimensional astronomical data sets. *2013 IEEE International Conference on Big Data*, 35–41. https://doi.org/10.1109/BigData.2013.6691731

Fan, X. (2006). Evolution of high-redshift quasars. *New Astronomy Reviews*, *50*(9), 665–671. https://doi.org/10.1016/j.newar.2006.06.077

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. (1996). The Sloan Digital Sky Survey Photometric System. *The Astronomical Journal*, *111*, 1748. https://doi.org/10.1086/117915

Greiner, J., Bolmer, J., Yates, R. M., Habouzit, M., Bañados, E., Afonso, P. M. J., & Schady, P. (2021). Quasar clustering at redshift 6. *Astronomy & Astrophysics*, *654*, A79. https://doi.org/10.1051/0004-6361/202140790

Hewett, P. C. (2010). *Improved redshifts for SDSS quasar spectra | Monthly Notices of the Royal Astronomical Society | Oxford Academic*. https://academic.oup.com/mnras/article/405/4/2302/1045471

Hickey, E., Creaner, O., Nolan, K., & O'Flynn, T. (2023). *Using machine learning to predict the correlation of spectra using SDSS magnitudes as an improvement to the Locus Algorithm—ScienceDirect*. https://www.sciencedirect.com/science/article/abs/pii/S1384107623000386

Kamath, C., Newsam, S., & Cantú-Paz, E. (2004). *Feature selection in scientific applications | Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. https://dl.acm.org/doi/abs/10.1145/1014052.1016915

Lin, Z. (2023). Classification of GALAXY, QSO, and STAR Based on KNN and PCA: *Proceedings of the 1st International Conference on Data Analysis and Machine Learning*, 54–60. https://doi.org/10.5220/0012814900003885

Peng, N., Zhang, Y., & Zhao, Y. (2013). A SVM-kNN method for quasar-star classification. *Science China Physics, Mechanics and Astronomy*, *56*(6), 1227–1234. https://doi.org/10.1007/s11433-013-5083-8

Richards, G. T., Nichol, R. C., Gray, A. G., & Brunner, R. J. (2004). *Efficient photometric selection of quasars from the SDSS. The Astrophysical Journal Supplement Series. 155(2)*(257.). https://doi.org/10.1086/425350

Solorio-Ramírez, J.-L., Jiménez-Cruz, R., Villuendas-Rey, Y., & Yáñez-Márquez, C. (2023). Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects. *Algorithms*, *16*(6). https://doi.org/10.3390/a16060293

Suwaid, M. A. H. B. M., Karim, M. 'Ilyas A. A., Hassan, R., & Aziz, A. A. (2025). Automated Classification of Celestial Objects Using Machine Learning. *International Journal on Perceptive and Cognitive Computing*, *11*(2), 22–41. https://doi.org/10.31436/ijpcc.v11i2.537

Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., LaBarbera, F. L., Capelato, H. V., Frago Campos Velho, H., Trevisan, M., & Ruiz, R. S. R. (2011). DECISION TREE CLASSIFIERS FOR STAR/GALAXY SEPARATION. *The Astronomical Journal*, *141*(6), 189. https://doi.org/10.1088/0004-6256/141/6/189

Vavilova, I. B., Dobrycheva, D. V., Vasylenko, M. Y., Elyiv, A. A., Melnyk, O. V., & Khramtsov, V. (2021). Machine learning technique for morphological classification of galaxies from the SDSS - I. Photometry-based approach. *Astronomy & Astrophysics*, *648*, A122. https://doi.org/10.1051/0004-6361/202038981

Warren, S. J., & Hewett, P. C. (1990). The detection of high-redshift quasars. *Reports on Progress in Physics*, *53*(8), 1095. https://doi.org/10.1088/0034-4885/53/8/003

York, D. G., Adelman, J., Anderson, Jr., J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., … Yasuda, N. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, *120*(3), 1579–1587. https://doi.org/10.1086/301513

Zheng, H., & Zhang, Y. (2008). Feature selection for high-dimensional data in astronomy. *Advances in Space Research*, *41*(12), 1960–1964. https://doi.org/10.1016/j.asr.2007.08.033