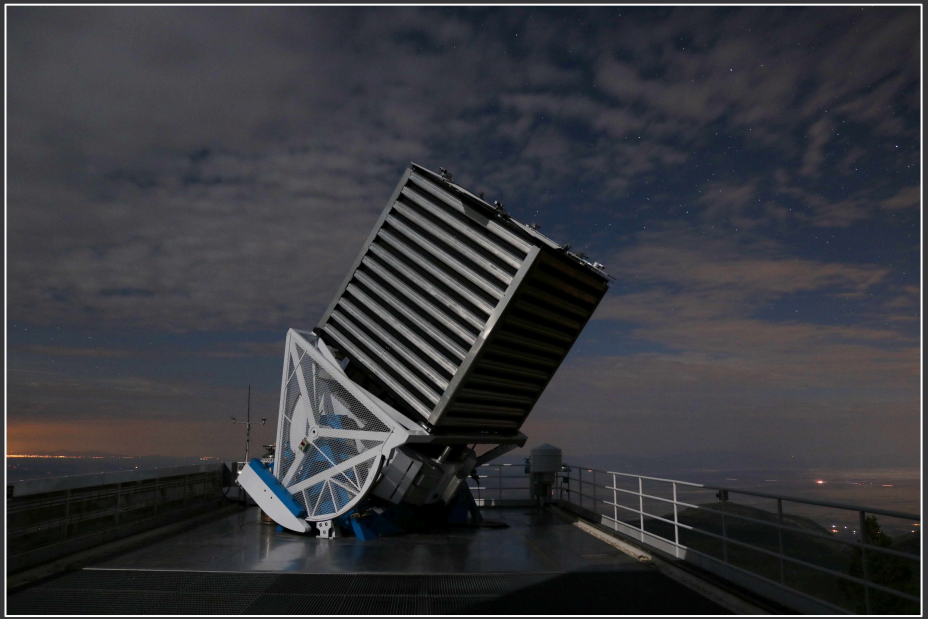


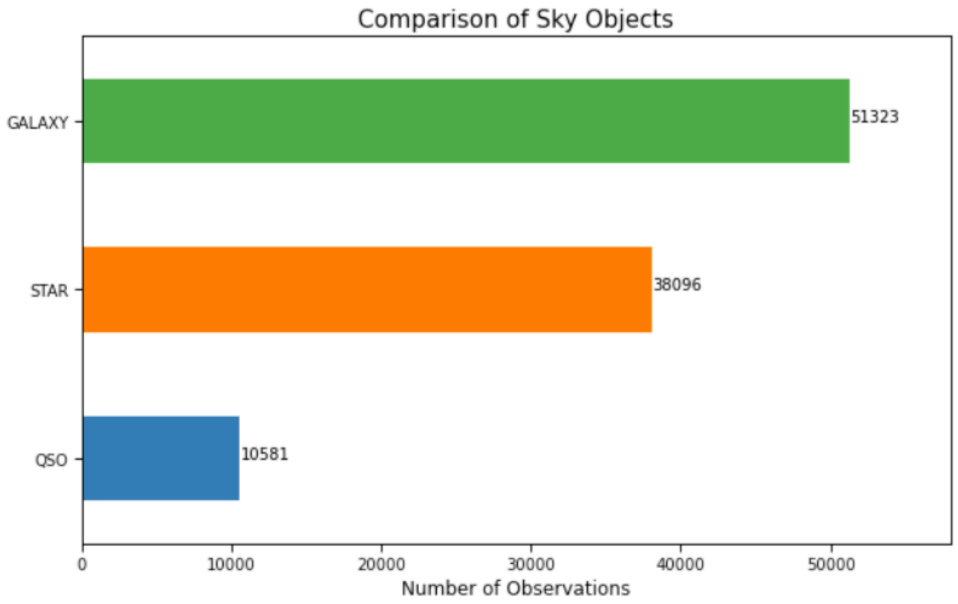
# SDSS Classification:

## Predicting the Nature of Celestial Bodies

Samuel Robbins

# Project Objective

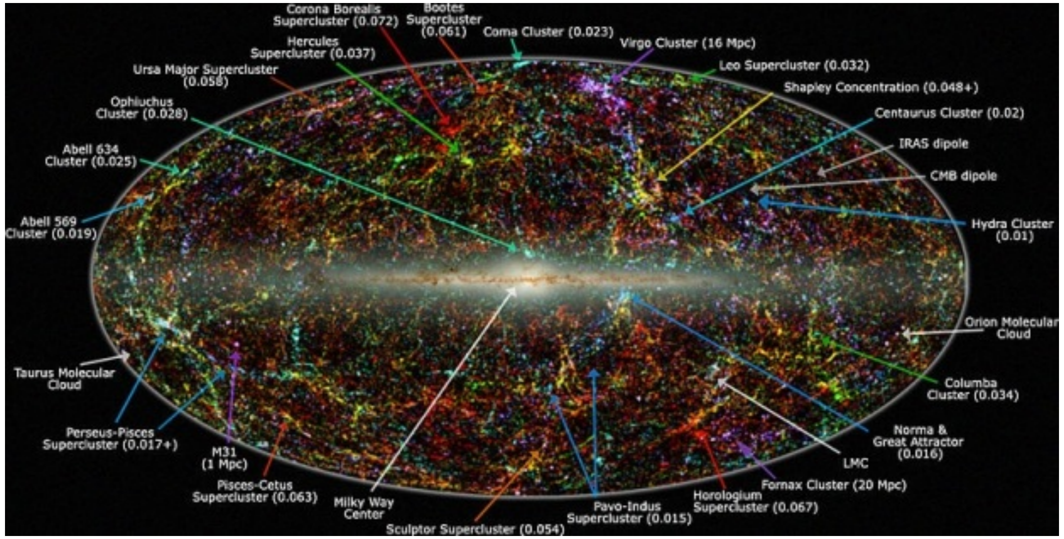
- Multi-class problem using spectral and imaging data from the Sloan Digital Sky Survey
- Predict the nature of celestial objects
- Evaluation metric = classification rate (accuracy)



# Data - SDSS Survey

- Sloan Digital Sky Survey – Data release 16
  - “The Sloan Digital Sky Survey has created the **most detailed three-dimensional maps of the Universe ever made**, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects.”

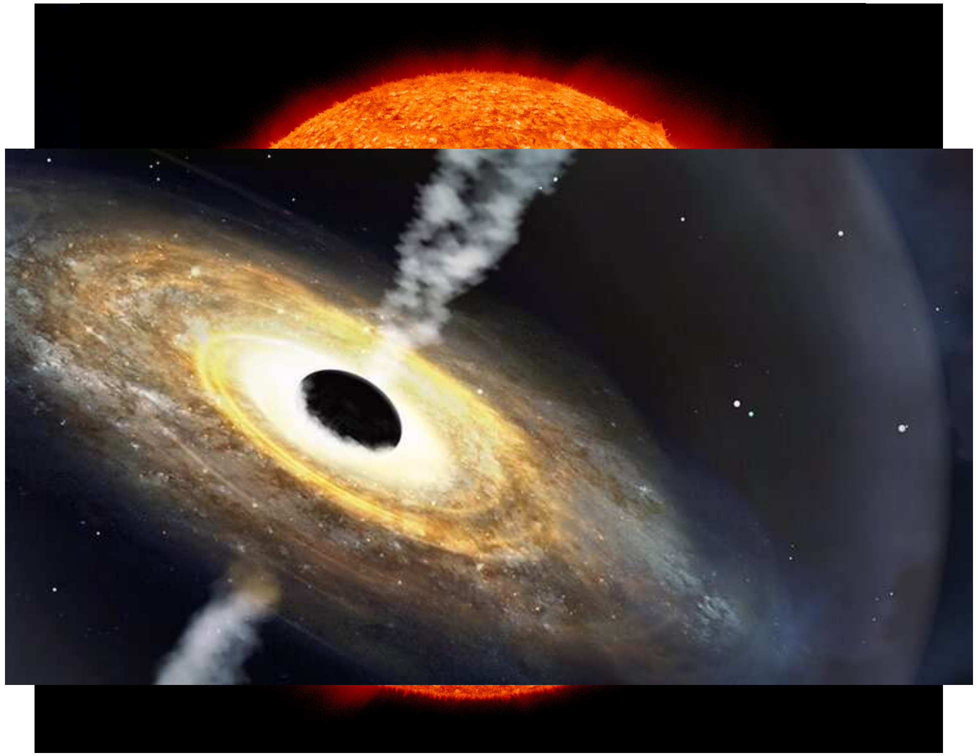
| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |


Corona Borealis Supercluster (0.072)
Bootes Supercluster (0.061)
Coma Cluster (0.023)
Virgo Cluster (16 Mpc)
Hercules Supercluster (0.037)
Leo Supercluster (0.032)
Ursa Major Supercluster (0.058)
Shapley Concentration (0.048+)
Ophiuchus Cluster (0.028)
Centaurus Cluster (0.02)
IRAS dipole
Abell 634 Cluster (0.025)
CMB dipole
Abell 569 Cluster (0.019)
Hydra Cluster (0.01)
Orion Molecular Cloud
Taurus Molecular Cloud
Columba Cluster (0.034)
Perseus-Pisces Supercluster (0.017+)
M31 (1 Mpc)
Norma & Great Attractor (0.016)
LMC
Fornax Cluster (20 Mpc)
Pisces-Cetus Supercluster (0.063)
Milky Way Center
Horologium Supercluster (0.067)
Pavo-Indus Supercluster (0.015)
Sculptor Supercluster (0.054)

# Data

## Classification Target – Class

- Galaxy
- Star
- Quasar Object

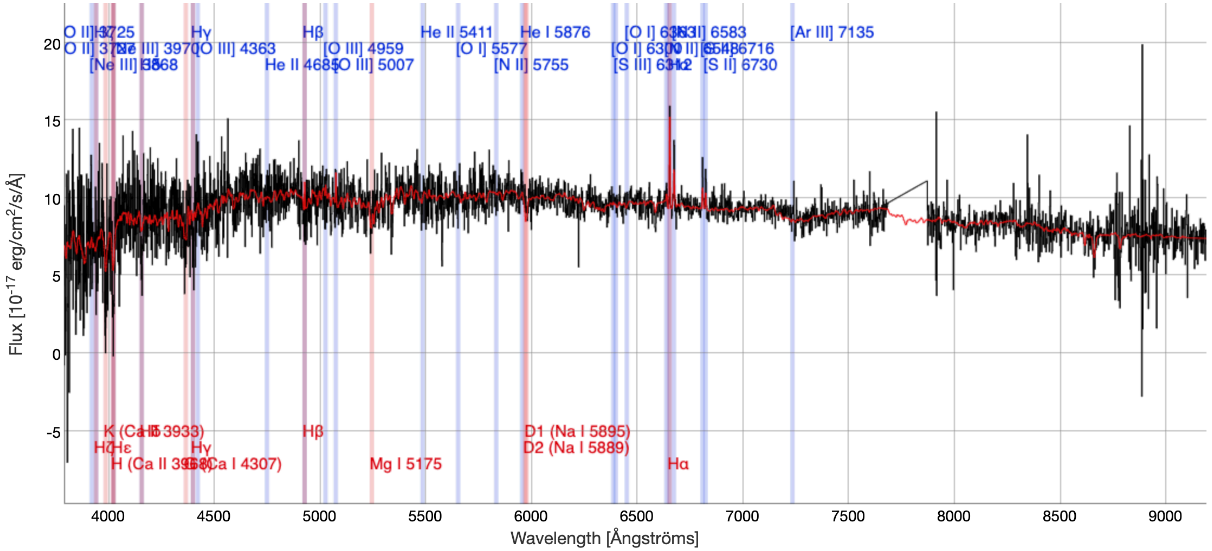


| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| **class** |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model



| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

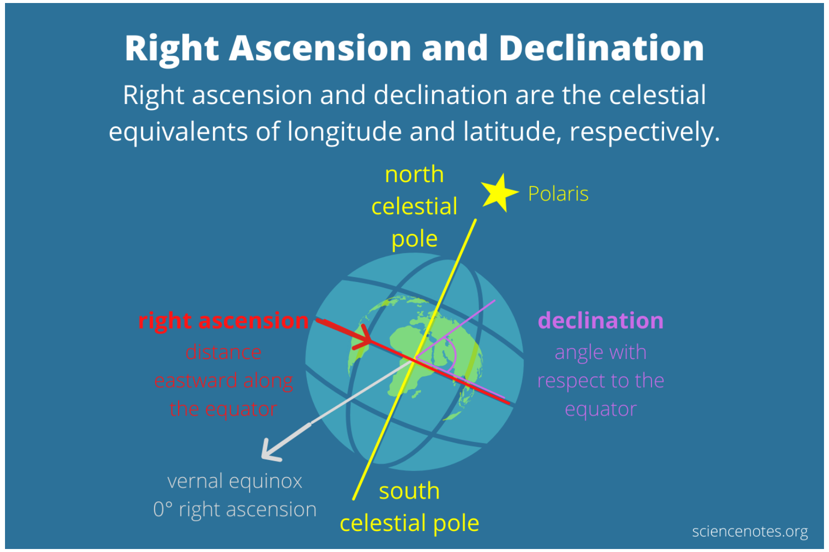
- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model
- Spectral Data
  - Object specific data

| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model
- Spectral Data
  - Object specific data
    - Right ascension/declination



**Right Ascension and Declination**

Right ascension and declination are the celestial equivalents of longitude and latitude, respectively.

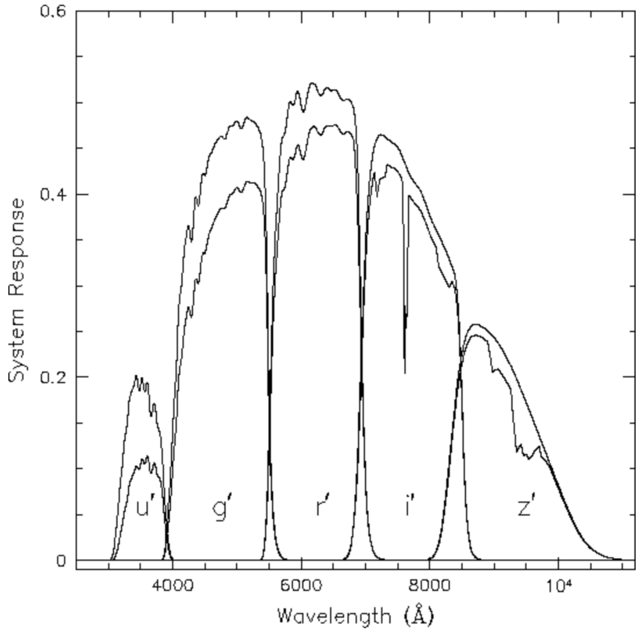
sciencenotes.org

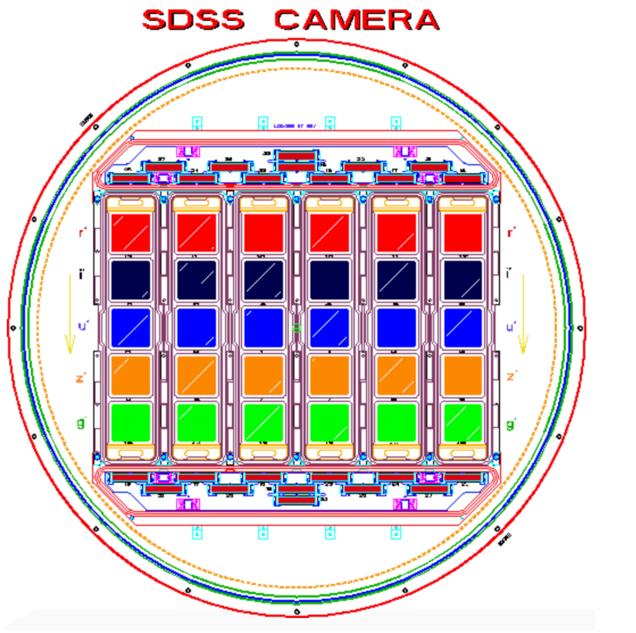
| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model
- Spectral Data
  - Object specific data
    - Right ascension/declination
    - 5-color photometric color system



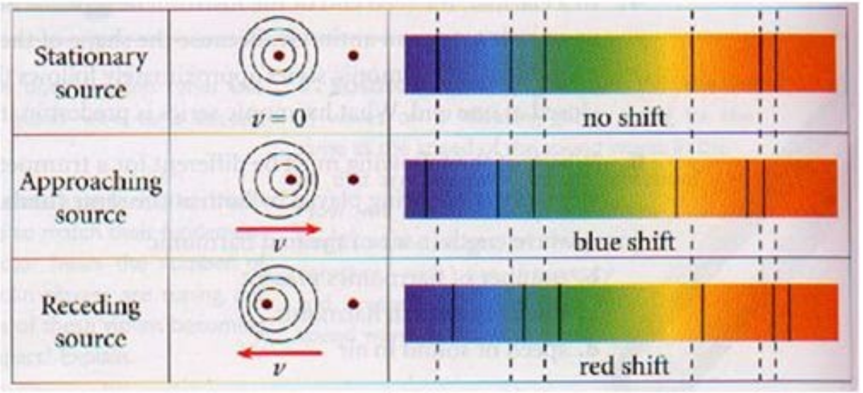


| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model
- Spectral Data
  - Object specific data
    - Right ascension/declination
    - 5-color photometric color system
    - Redshift



Measuring the relative velocities of stars by the Doppler shift.

| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Data

## Classification Features

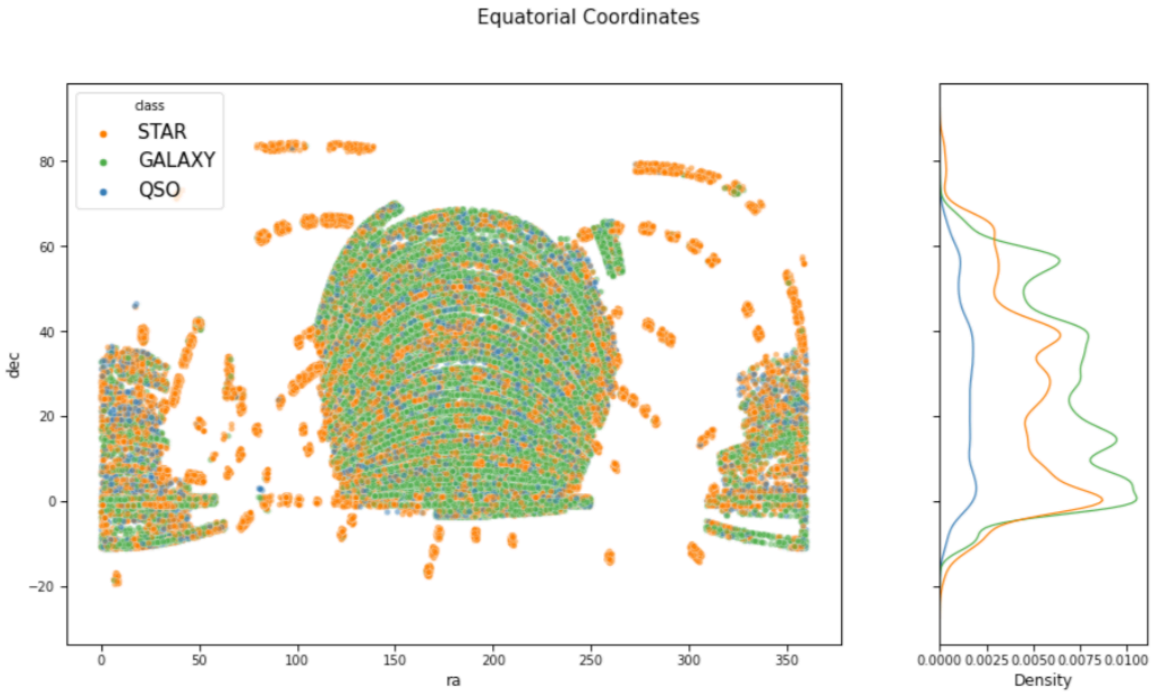
- Image/Identification Data
  - Ex. run, rerun, and camcol describe a field within an image
  - NOT used in classification model
- Spectral Data
  - Object specific data
    - 5-color photometric color system
    - Redshift
  - Used in classification model

## Classification Target – Class

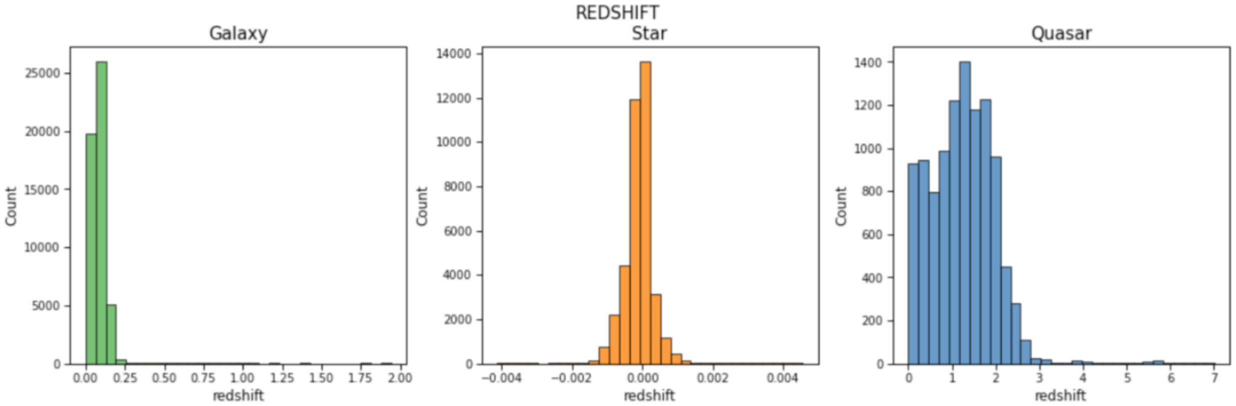
- Galaxy
- Star
- Quasar Object

| FEATURES |
|---|
| Objid |
| ra |
| dec |
| u-band |
| g-band |
| r-band |
| i-band |
| z-band |
| run |
| rerun |
| camcol |
| field |
| specobjid |
| class |
| redshift |
| plate |
| mjd |
| fiberid |

# Exploratory Analysis

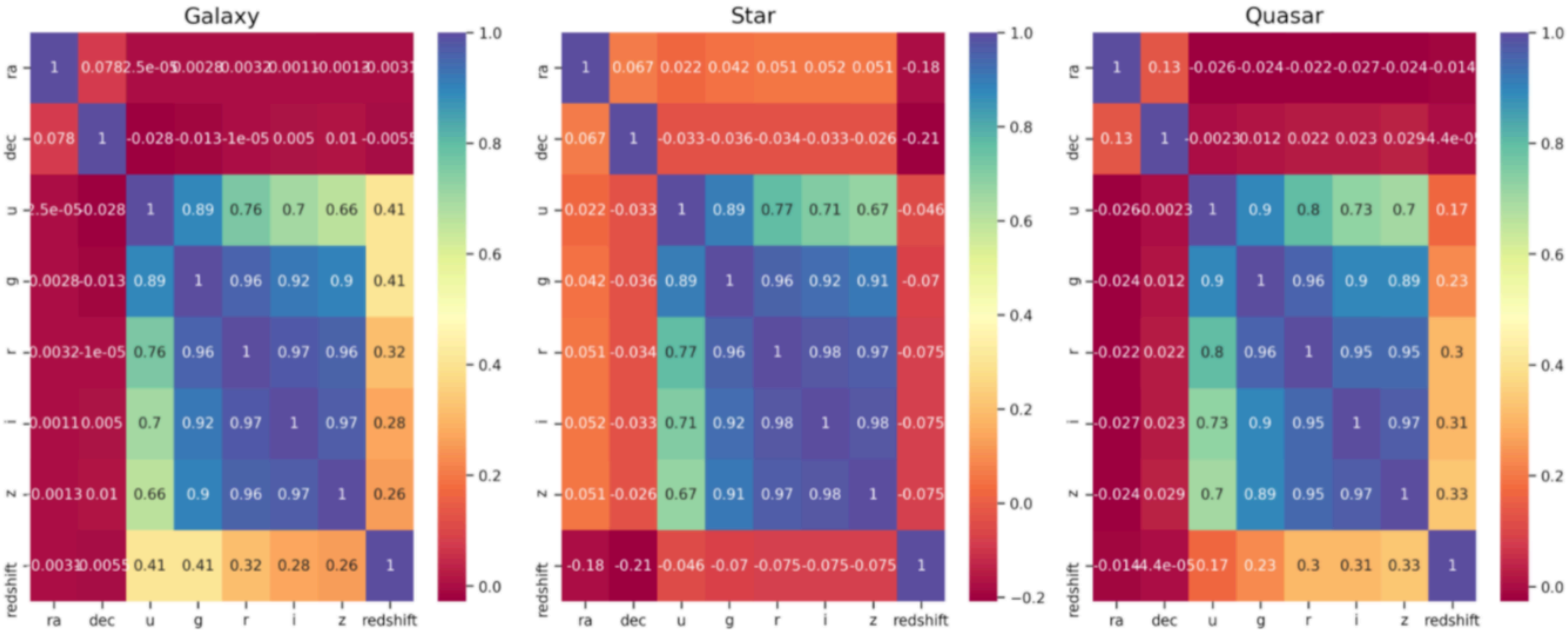


- Is position in the sky a distinct feature for the three class types?



- Redshift is the most characteristic feature for the different classes.

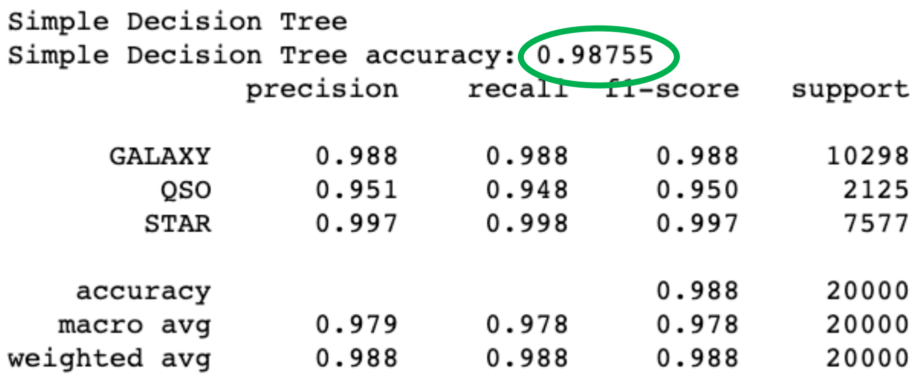
# Exploratory Analysis



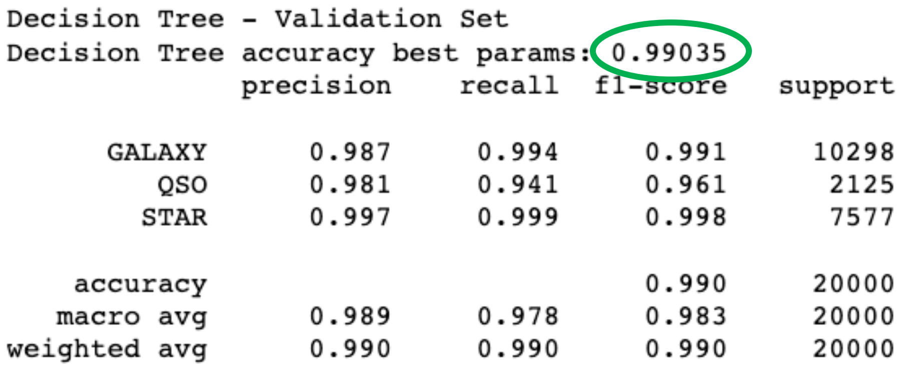
Photometric color data is the most correlated within the dataset.

# Algorithms and Final Model

- Logistic Regression
- K-Nearest Neighbors
- Decision Trees – best model
- Random Forests
  - *Computationally expensive to tune hyperparameters*
- Naïve Bayes



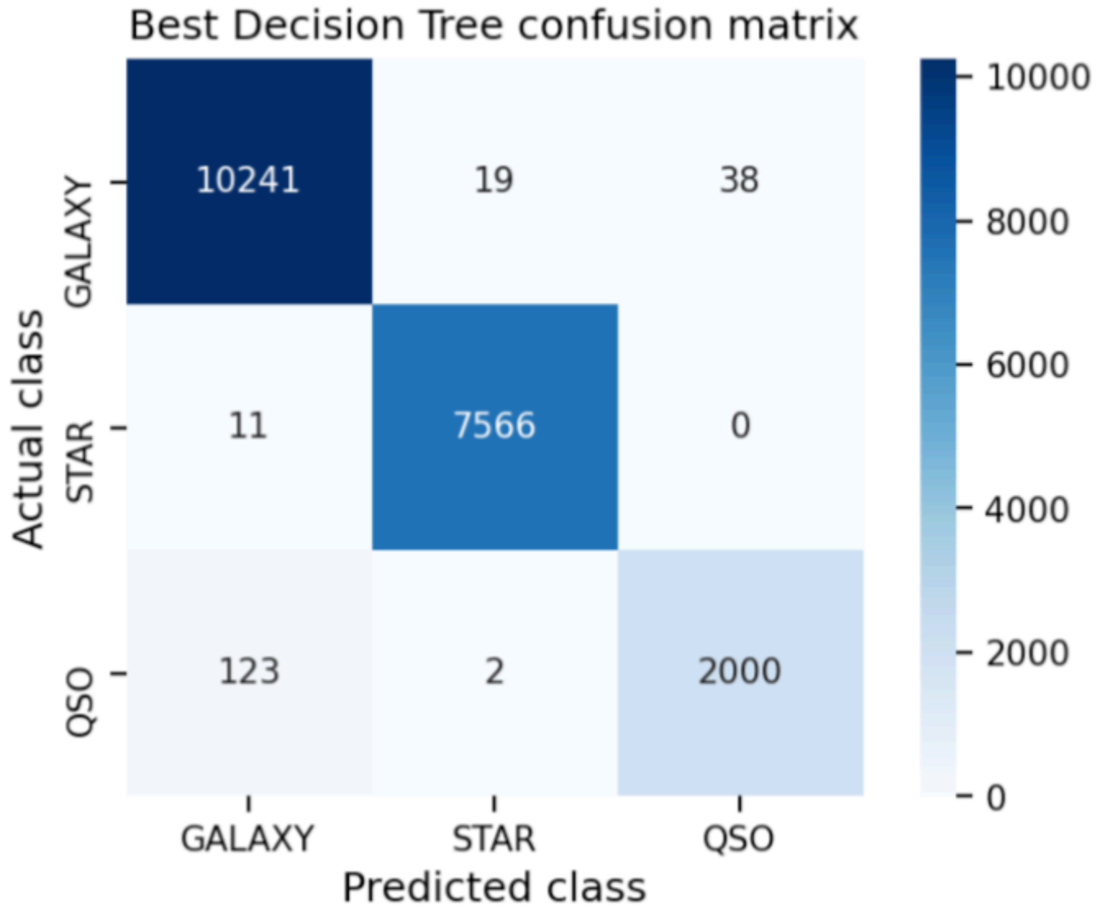
```
Simple Decision Tree
Simple Decision Tree accuracy: 0.98755
              precision    recall  f1-score   support

      GALAXY      0.988     0.988     0.988     10298
         QSO      0.951     0.948     0.950      2125
        STAR      0.997     0.998     0.997      7577

    accuracy                          0.988     20000
   macro avg      0.979     0.978     0.978     20000
weighted avg      0.988     0.988     0.988     20000
```



```
Decision Tree - Validation Set
Decision Tree accuracy best params: 0.99035
              precision    recall  f1-score   support

      GALAXY      0.987     0.994     0.991     10298
         QSO      0.981     0.941     0.961      2125
        STAR      0.997     0.999     0.998      7577

    accuracy                          0.990     20000
   macro avg      0.989     0.978     0.983     20000
weighted avg      0.990     0.990     0.990     20000
```

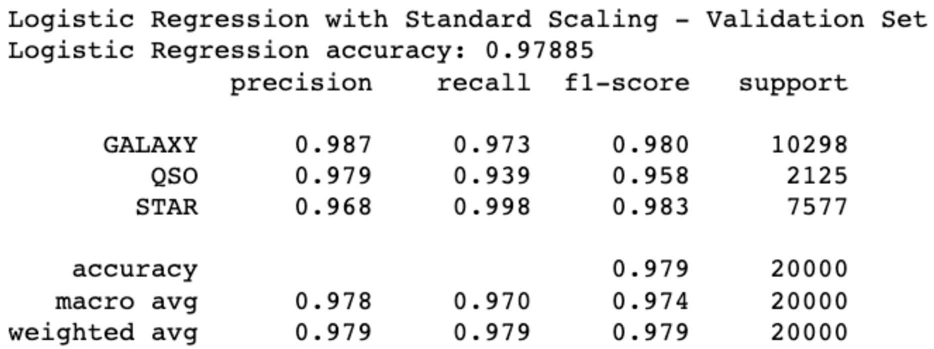
# Algorithms and Final Model

- Logistic Regression
- K-Nearest Neighbors
- Decision Trees – best model
- Random Forests
  - *Computationally expensive to tune hyperparameters*
- Naïve Bayes

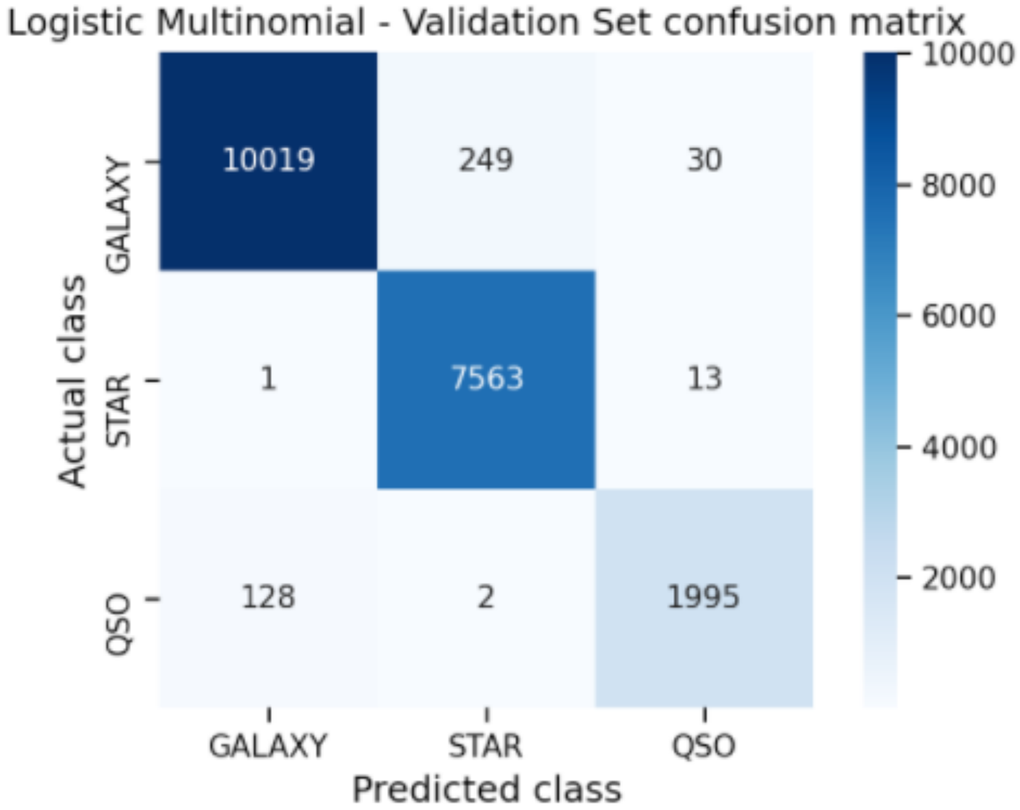


# Appendix

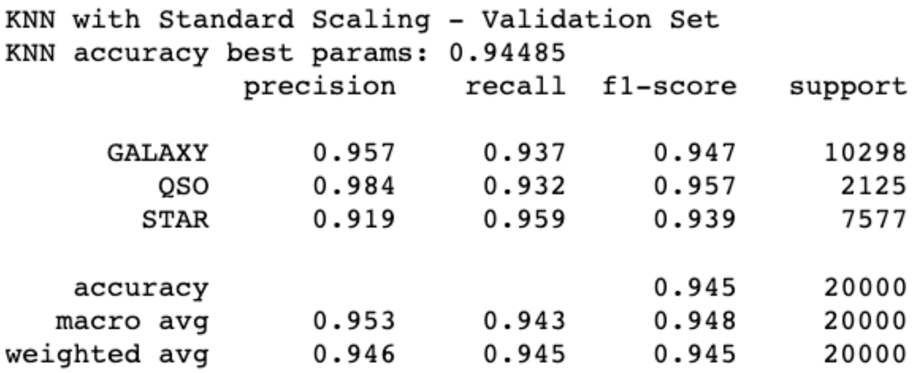
# Logistic Regression



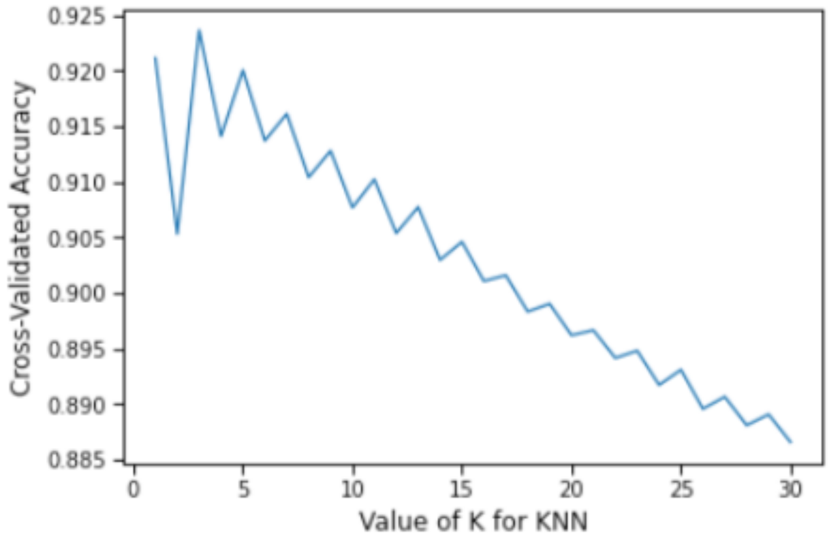
```
Logistic Regression with Standard Scaling - Validation Set
Logistic Regression accuracy: 0.97885
              precision    recall  f1-score   support

      GALAXY      0.987     0.973     0.980     10298
         QSO      0.979     0.939     0.958      2125
        STAR      0.968     0.998     0.983      7577

    accuracy                          0.979     20000
   macro avg      0.978     0.970     0.974     20000
weighted avg      0.979     0.979     0.979     20000
```

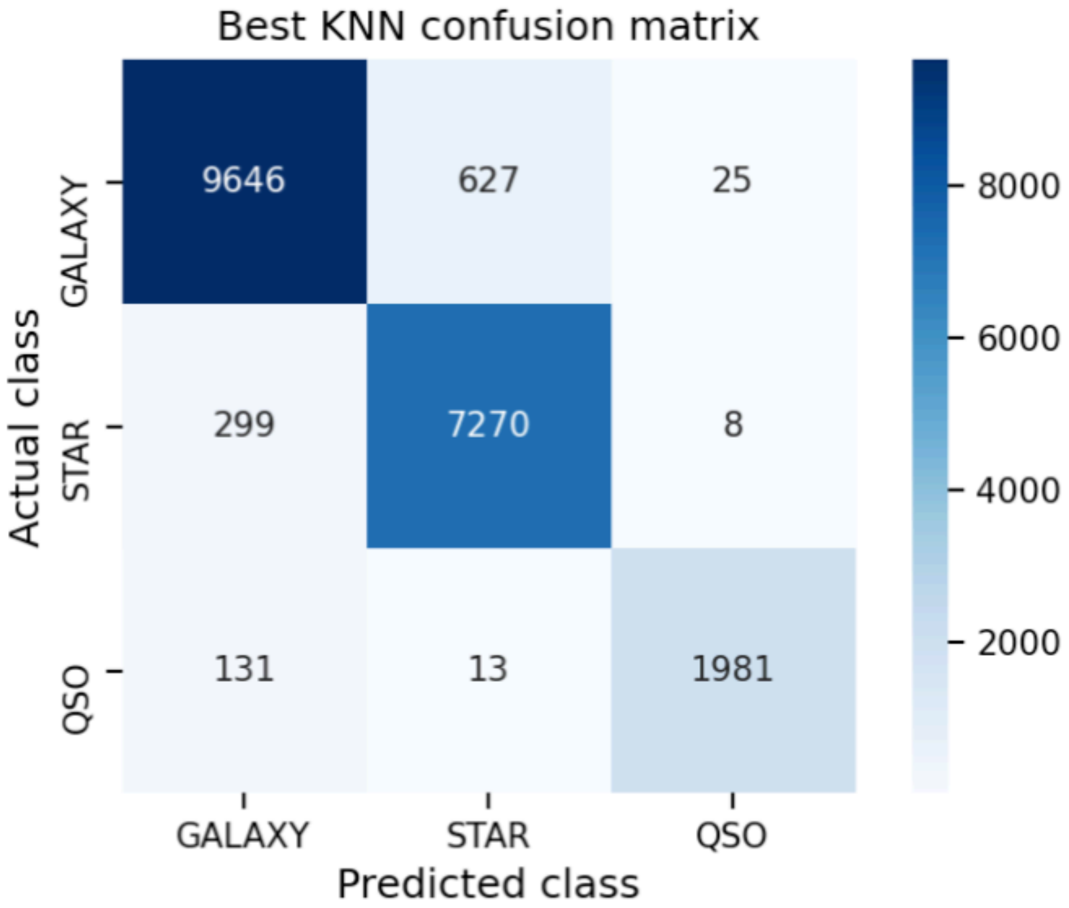


# K-Nearest Neighbor

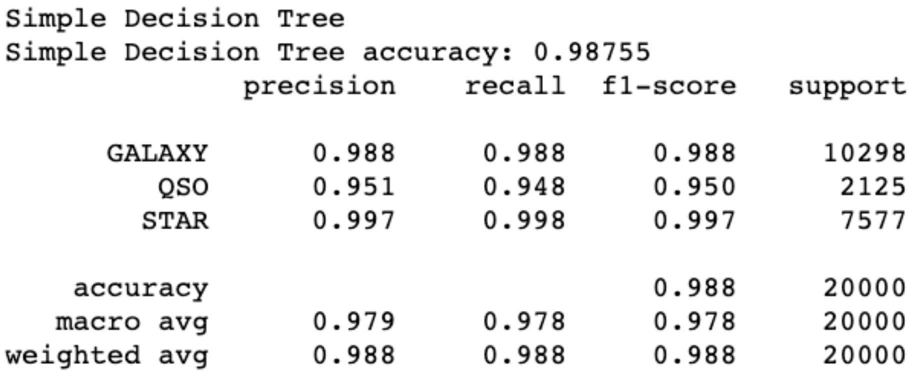


```
KNN with Standard Scaling - Validation Set
KNN accuracy best params: 0.94485
              precision    recall  f1-score   support

      GALAXY      0.957     0.937     0.947     10298
         QSO      0.984     0.932     0.957      2125
        STAR      0.919     0.959     0.939      7577

    accuracy                          0.945     20000
   macro avg      0.953     0.943     0.948     20000
weighted avg      0.946     0.945     0.945     20000
```

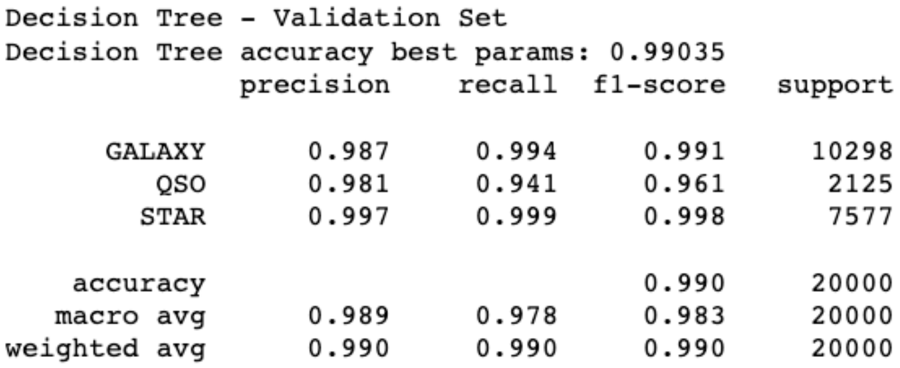




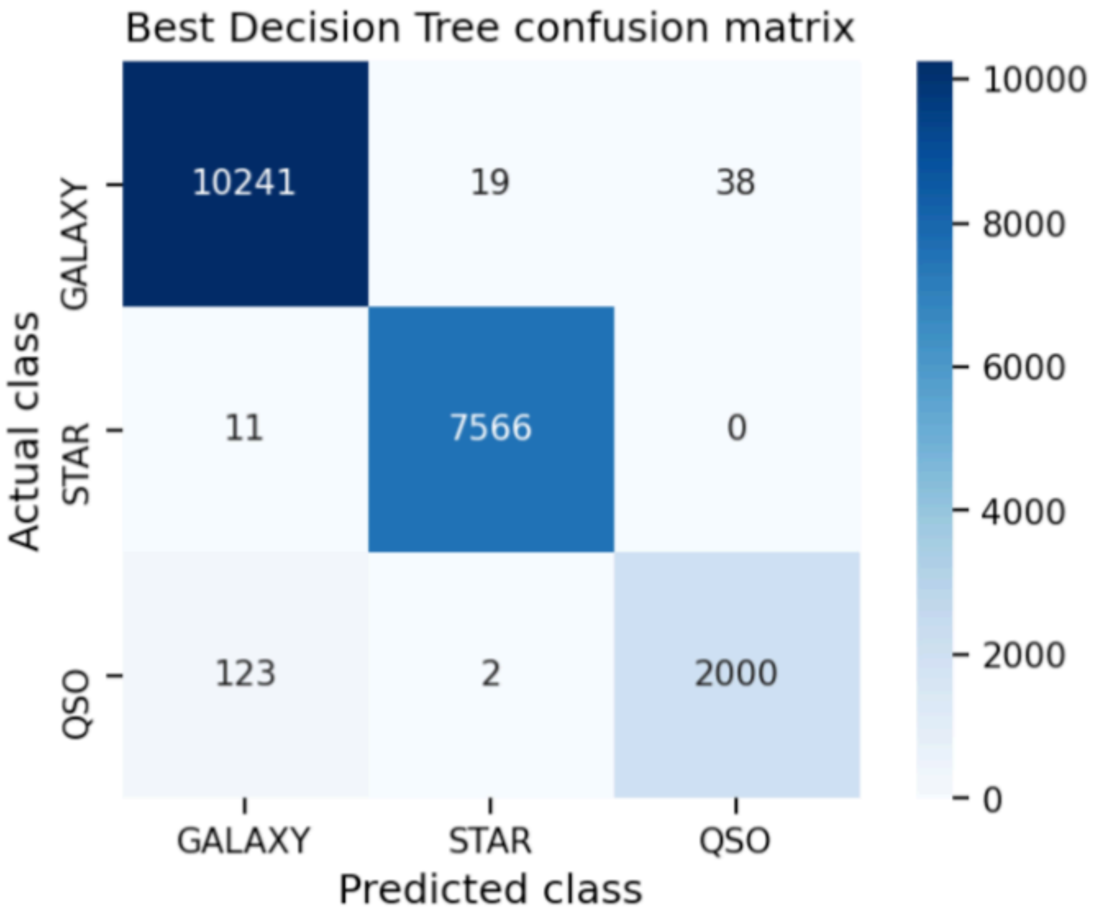
# Decision Tree



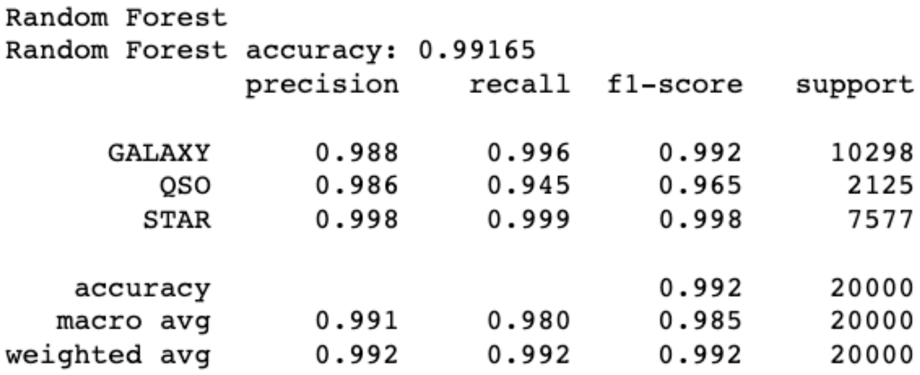
```
Simple Decision Tree
Simple Decision Tree accuracy: 0.98755
              precision    recall  f1-score   support

      GALAXY      0.988     0.988     0.988     10298
         QSO      0.951     0.948     0.950      2125
        STAR      0.997     0.998     0.997      7577

    accuracy                          0.988     20000
   macro avg      0.979     0.978     0.978     20000
weighted avg      0.988     0.988     0.988     20000
```



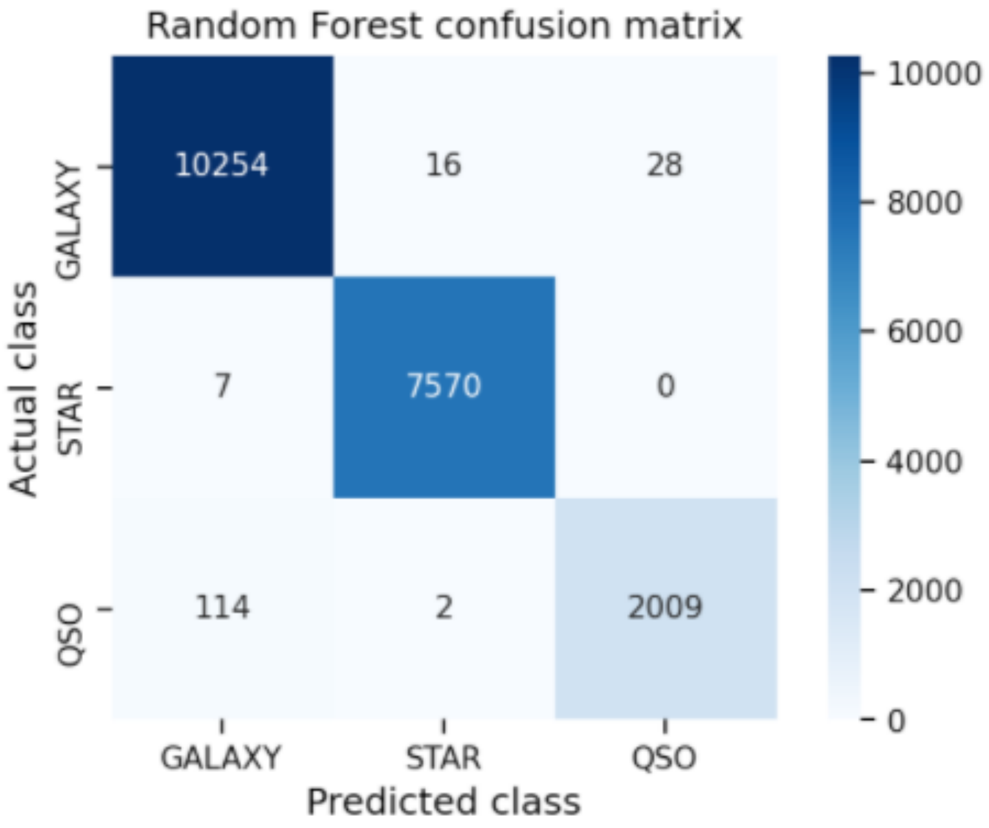
```
Decision Tree - Validation Set
Decision Tree accuracy best params: 0.99035
              precision    recall  f1-score   support

      GALAXY      0.987     0.994     0.991     10298
         QSO      0.981     0.941     0.961      2125
        STAR      0.997     0.999     0.998      7577

    accuracy                          0.990     20000
   macro avg      0.989     0.978     0.983     20000
weighted avg      0.990     0.990     0.990     20000
```



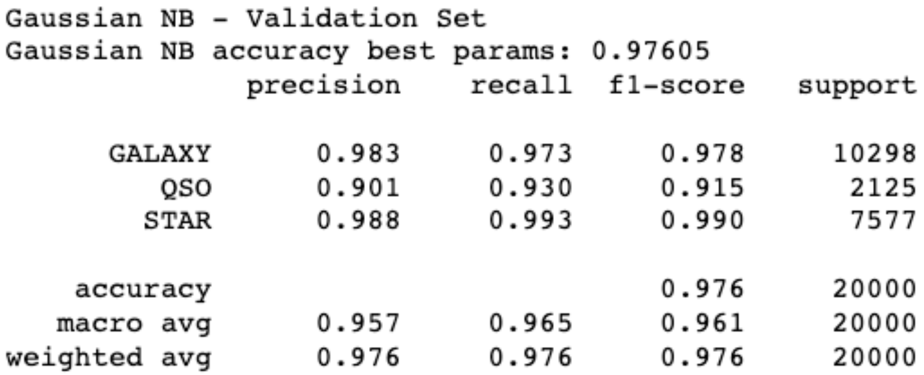
# Random Forests



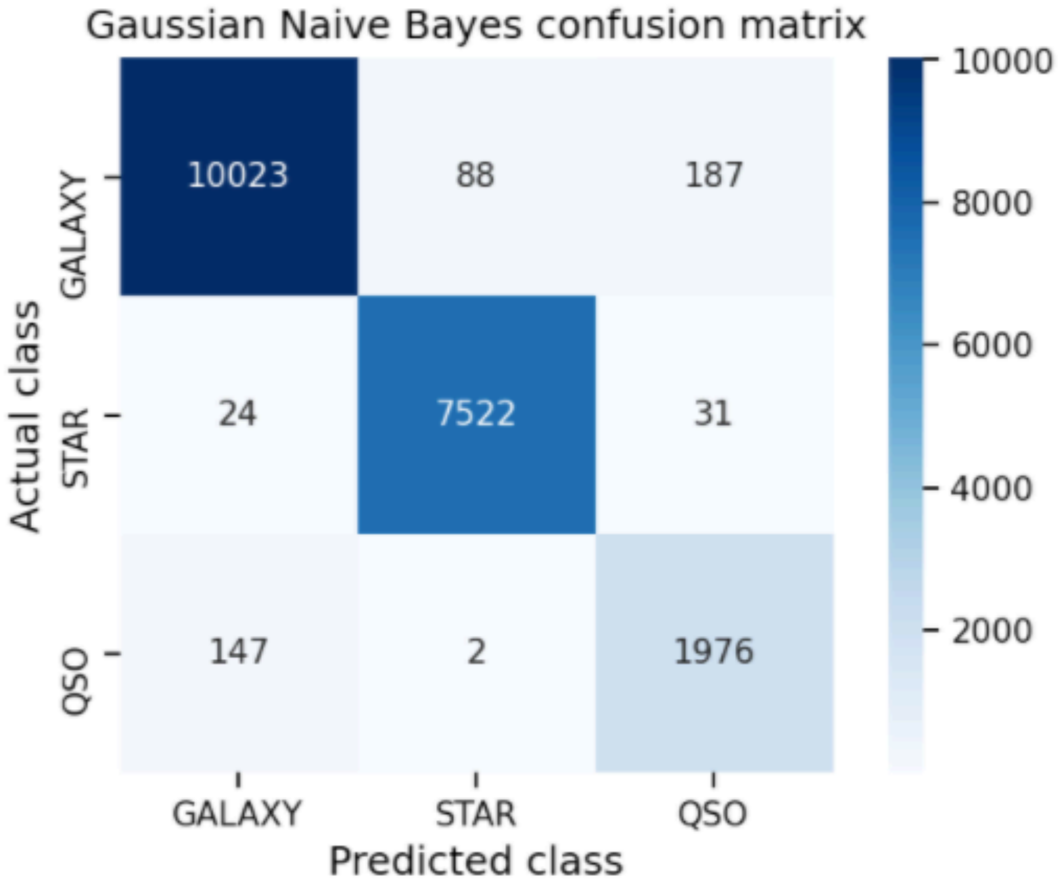
```
Random Forest
Random Forest accuracy: 0.99165
              precision    recall  f1-score   support

      GALAXY      0.988     0.996     0.992     10298
         QSO      0.986     0.945     0.965      2125
        STAR      0.998     0.999     0.998      7577

    accuracy                          0.992     20000
   macro avg      0.991     0.980     0.985     20000
weighted avg      0.992     0.992     0.992     20000
```



# Naïve Bayes



```
Gaussian NB - Validation Set
Gaussian NB accuracy best params: 0.97605
              precision    recall  f1-score   support

      GALAXY      0.983     0.973     0.978     10298
         QSO      0.901     0.930     0.915      2125
        STAR      0.988     0.993     0.990      7577

    accuracy                          0.976     20000
   macro avg      0.957     0.965     0.961     20000
weighted avg      0.976     0.976     0.976     20000
```



Gaussian Naive Bayes confusion matrix
Actual class
GALAXY
STAR
QSO
GALAXY
STAR
QSO
Predicted class