

# CLASSIFYING COSMIC POINT-SOURCES OF LIGHT WITH MACHINE LEARNING

Jenica Andersen

April 20, 2022

Metis DSML, Classification Module



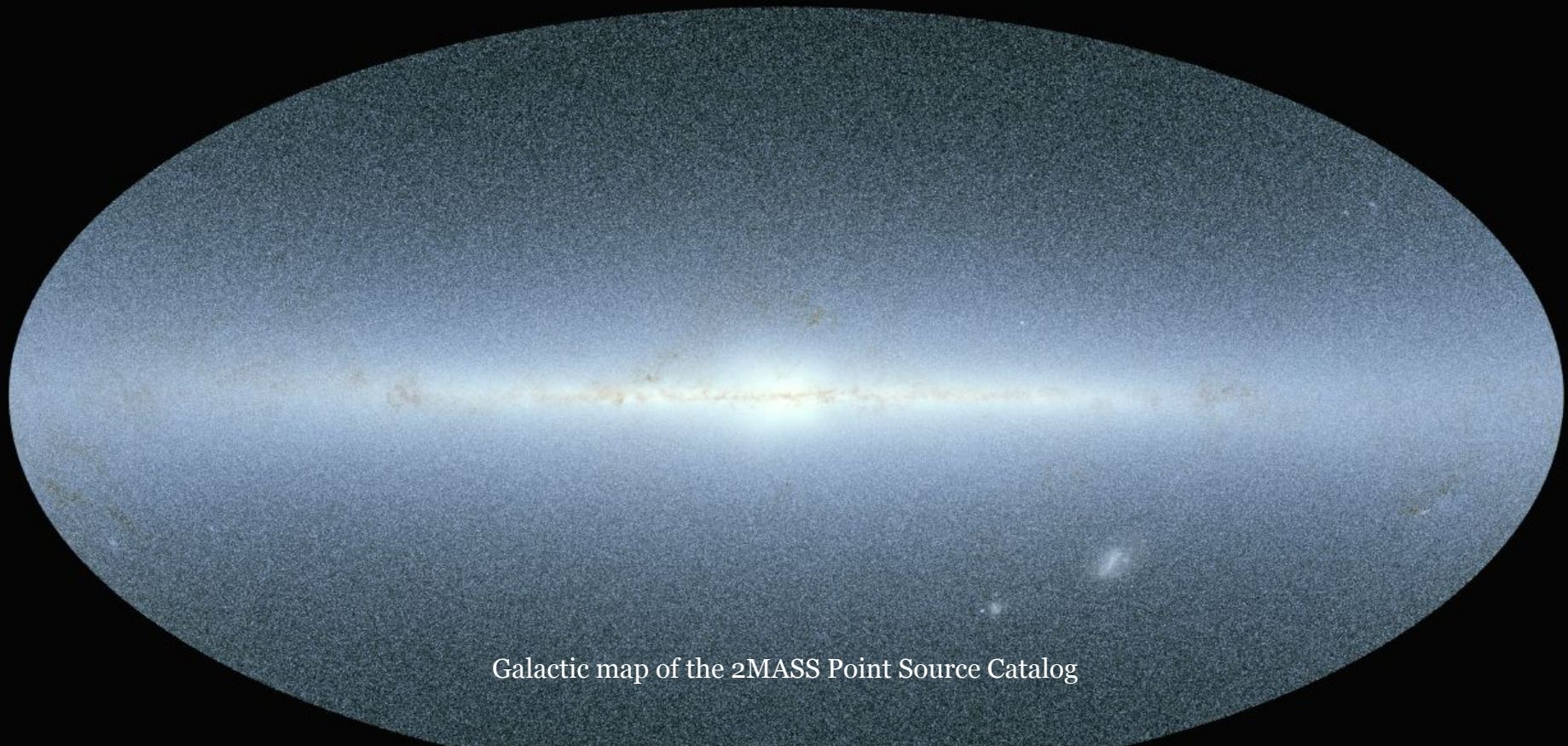
**SDSS**



**Los Alamos**  
NATIONAL LABORATORY



# PANORAMIC VIEW OF THE SKY



Galactic map of the 2MASS Point Source Catalog



# PANORAMIC VIEW OF THE SKY

**Unfeasible to classify all of these objects manually**

Galactic map of the 2MASS Point Source Catalog

# WHY MAP THE SKY?



## MAKE DISCOVERIES

Think cartography when Earth was uncharted



## ORIGINS OF THE UNIVERSE

Understand cosmic evolution and the history of where we all came from



## DARK MATTER

Map what's seen, to better understand what's unseen

# LIGHT POINT-SOURCES

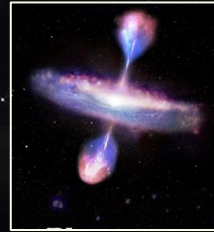
STARS



GALAXIES



QUASARS





# THE DATA

## KAGGLE STELLAR CLASSIFICATION DATASET:

- **100,000 “sources”**
  - **17 features**
    - **(8-10 “of interest”, all continuous)**
  - **Target: “class”**
    - **18% Quasar**
    - **60% Galaxy**
    - **22% Stars**
- (No class imbalance handling)

# FEATURES

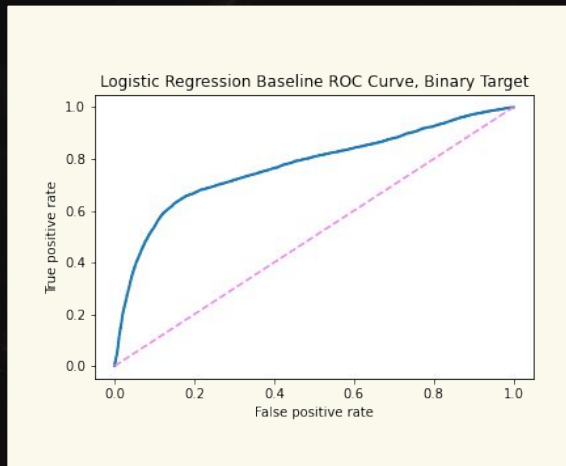
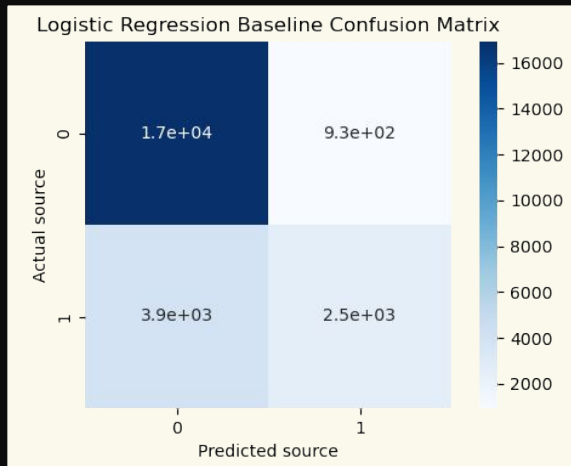
## Features Used:

- **Location: Coordinates** on celestial sphere, **Field #**
- Photometric data ("**brightness**"): u, g, r, i, z
- **Redshift** data: increase in wavelength

## Features NOT Used:

- Object, Equipment and Run ID's
- Date



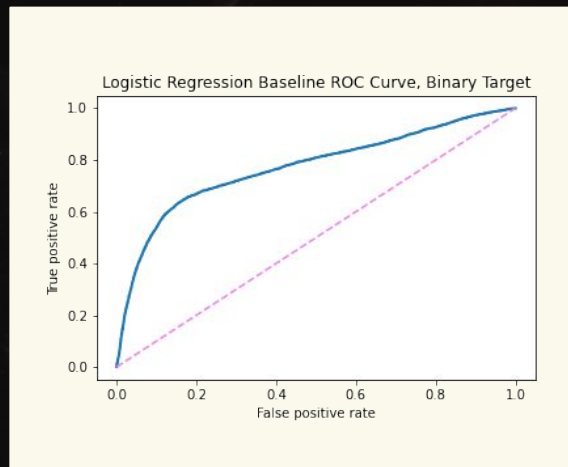
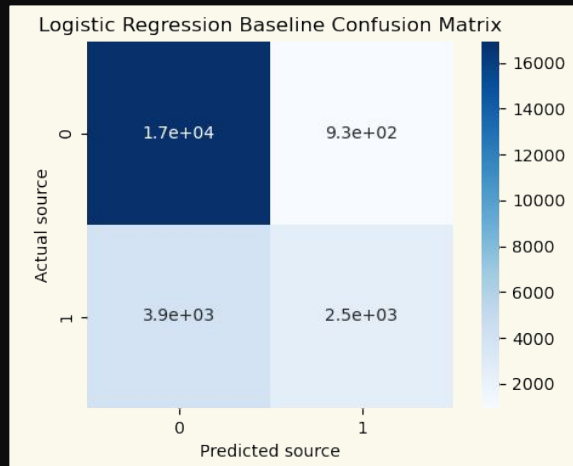


# EDA: BASELINE BINARY LOGISTIC REGRESSION

Positive Class: "Star"

ROC AUC Score = 0.77





# EDA: BASELINE BINARY LOGISTIC REGRESSION

Positive Class: "Star"

ROC AUC Score = 0.77

**4,830 sources misclassified!**

# MODELS & ALGORITHMS



KNN

PIPELINE

Organized models,  
preprocessing & parameters



LOGISTIC REGRESSION

RANDOMIZED-  
& GRIDSEARCH CV  
**Random-** was faster, honed  
by **Grid-**. 30% test data



RANDOM FOREST

SCALED DATA

MinMaxScaler and  
Standard Scaler



# METRICS

## ROC AUC

tells how much the model  
is capable of distinguishing  
between classes

ROC

R

## RECALL

indicates how good the  
classifier is at minimizing  
false negatives

## PRECISION

indicating how good the  
classifier is at identifying  
true positives

P

F1

## F1 SCORE

an overall performance  
metric



# METRICS

Used for GridSearchCV

## ROC AUC

tells how much the model  
is capable of distinguishing  
between classes

ROC

R

## RECALL

indicates how good the  
classifier is at minimising  
false negatives.

## PRECISION

indicating how good the  
classifier is at identifying  
true positives

P

F1

## F1 SCORE

an overall performance  
metric

# METRICS

Used for GridSearchCV

## ROC AUC

tells how much the model is capable of distinguishing between classes

ROC

## PRECISION

indicating how good the classifier is at identifying true positives

P

Also key

## RECALL

indicates how good the classifier is at minimising false negatives.

R

## F1 SCORE

an overall performance metric

F1

# RECALL: COST OF FALSE NEGATIVES > FALSE POSITIVES (IN BINARY MODEL)



- **Nearby stars** are brighter → prone to misclassification
- Strong interest in nearby **extra solar planets** (exploring for life, richer scientific results)



- **Fewer stars** than galaxies (single misclassification represents a larger fraction)
- Potential for **compounded mistakes**, far-reaching and unknown



# VALIDATION RESULTS

K-NEAREST  
NEIGHBOR

0.9900

LOGISTIC  
REGRESSION

0.9969

RANDOM  
FOREST

>0.999

# RESULTS: VALIDATION MODEL SCORES

K-NEAREST  
NEIGHBOR 0.9900

LOGISTIC  
REGRESSION 0.9969

RANDOM  
FOREST >0.999



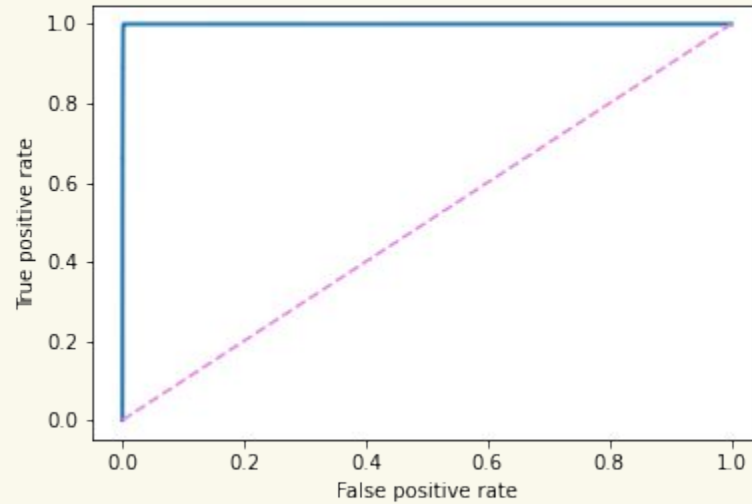
# RESULTS: FINAL ROC AUC SCORES (BINARY CLASS ONLY)

VALIDATION  
SCORE **>0.999**

TEST DATA  
SCORE **0.9983**

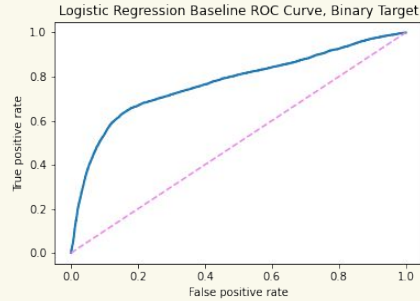


ROC curve for Random Forest, Star vs Galaxy Point Source Prediction

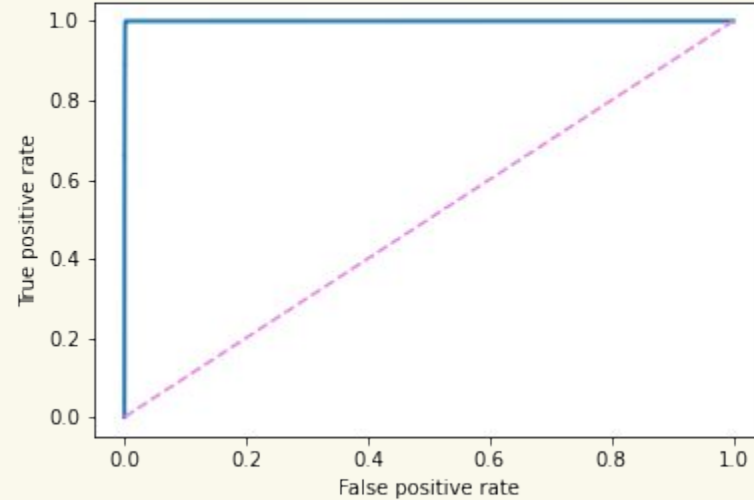


# FINAL RANDOM FOREST MODEL (BINARY CLASS ONLY)

## Baseline model

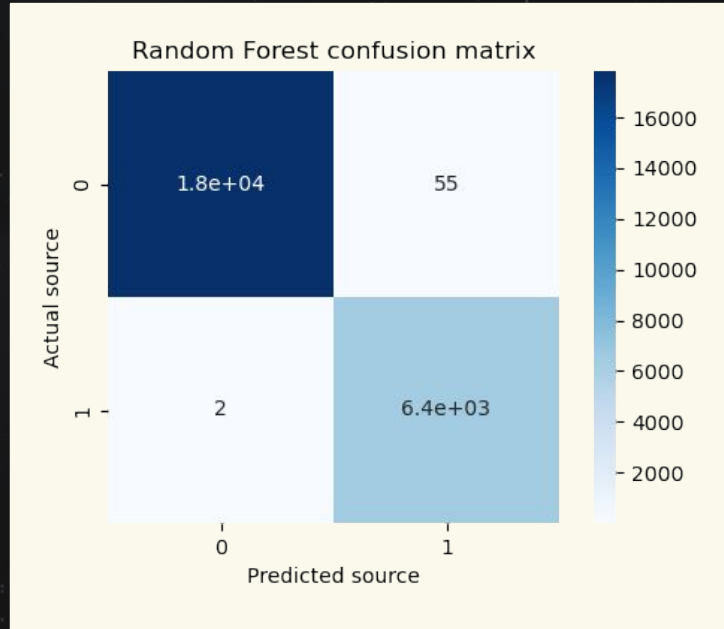


ROC curve for Random Forest, Star vs Galaxy Point Source Prediction

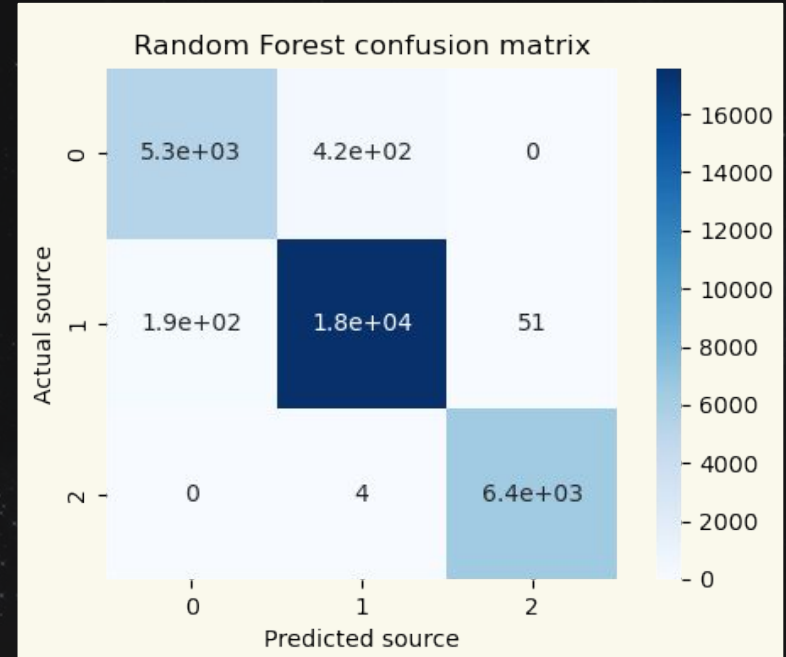


# FINAL RANDOM FOREST MODEL (BINARY CLASS ONLY)

# BINARY

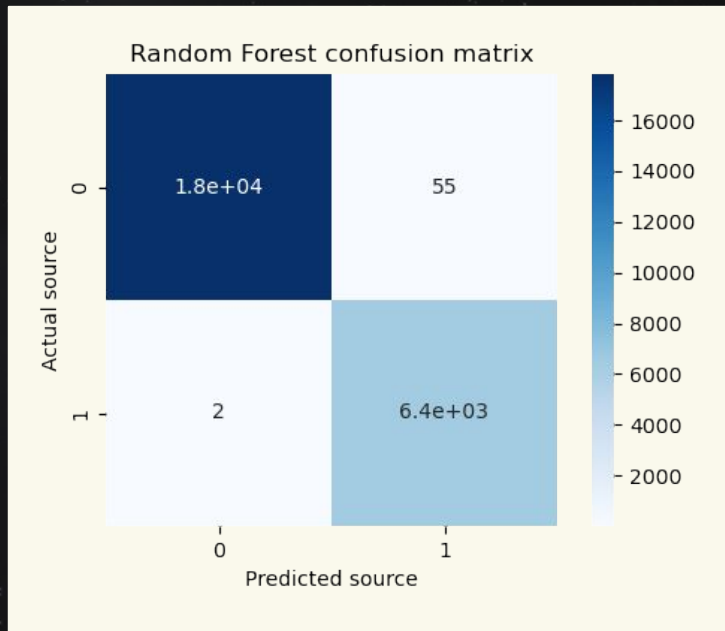


# MULTICLASS



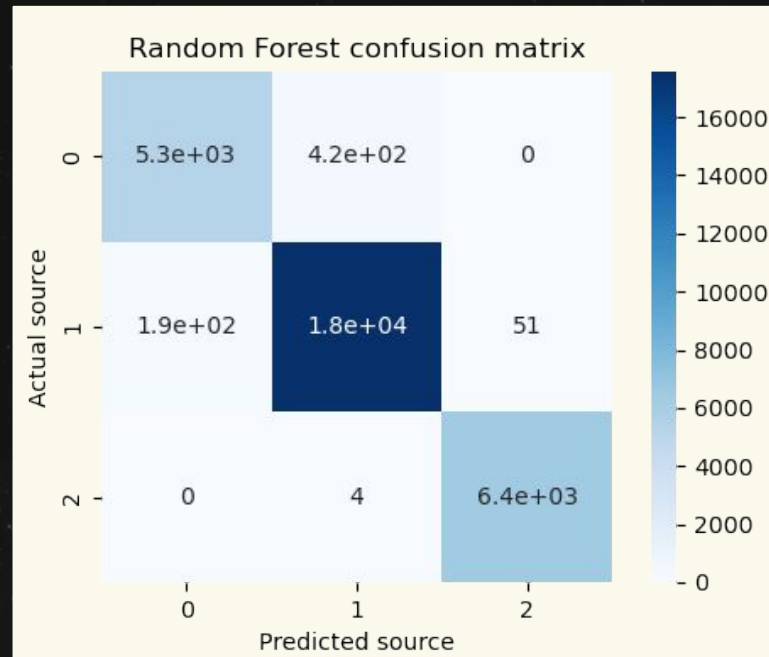


# BINARY



**57 sources misclassified!**  
**Recall = >0.9997**

# MULTICLASS



**665 sources misclassified!**  
**Recall = 0.9708**

# LOOKING AHEAD



TRAIN AND TUNE MULTICLASS  
MODELS



INSPECT MISCLASSIFICATIONS



TRY STACKING OR ITERATIVE  
CLASSIFIERS

# SUMMARY



## MAPPING THE COSMOS IS IMPORTANT

Don't know the Full Impact or All Applications Yet



## RANDOM FOREST WAS BEST PERFORMER

ROC Score of 0.9983 vs EDA Logistic  
Regression Score 0.77



## MODEL CAN BE IMPROVED

Tune for Multiclass and try stacking or voting  
classifiers



# REFERENCES



- Sloan Digit Sky Survey <https://www.sdss.org/dr17/>
- Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra A. O. Clarke, A. M. M. Scaife, R. Greenhalgh and V. Griguta A&A, 639 (2020) A84 DOI: <https://doi.org/10.1051/0004-6361/201936770>



# THANKS!

Does anyone have any questions?

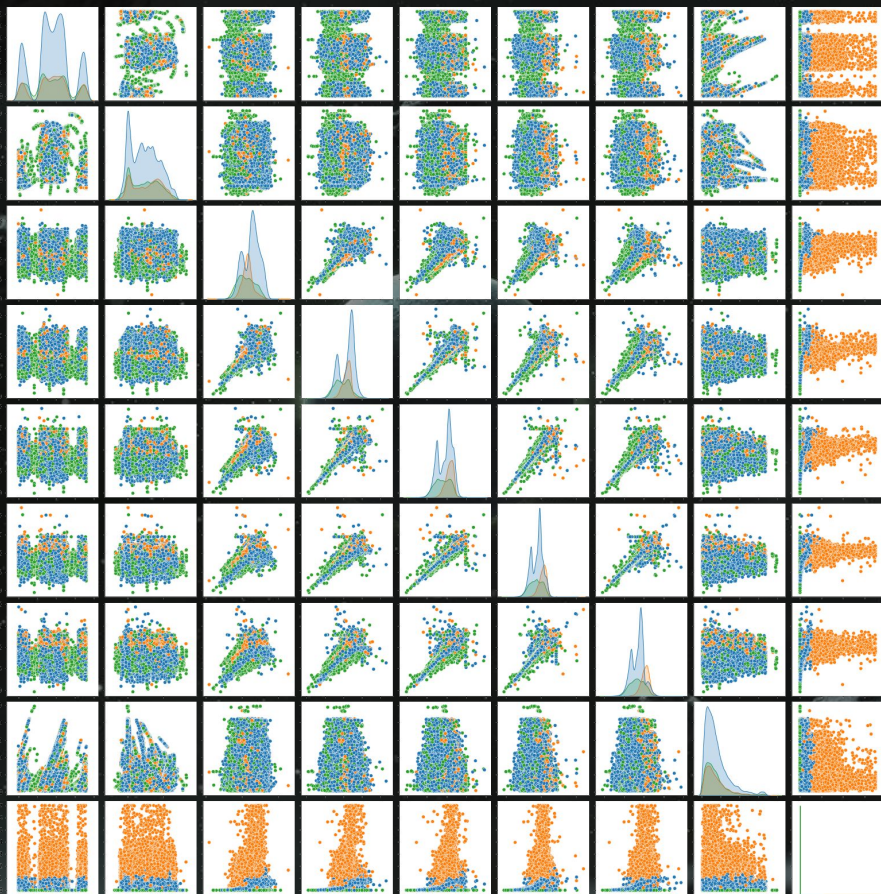


**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

INTRODUCTION - METHOD - RESULTS - CONCLUSION - APPENDIX

# FINAL RANDOM FOREST PARAMETERS (BINARY CLASS ONLY)

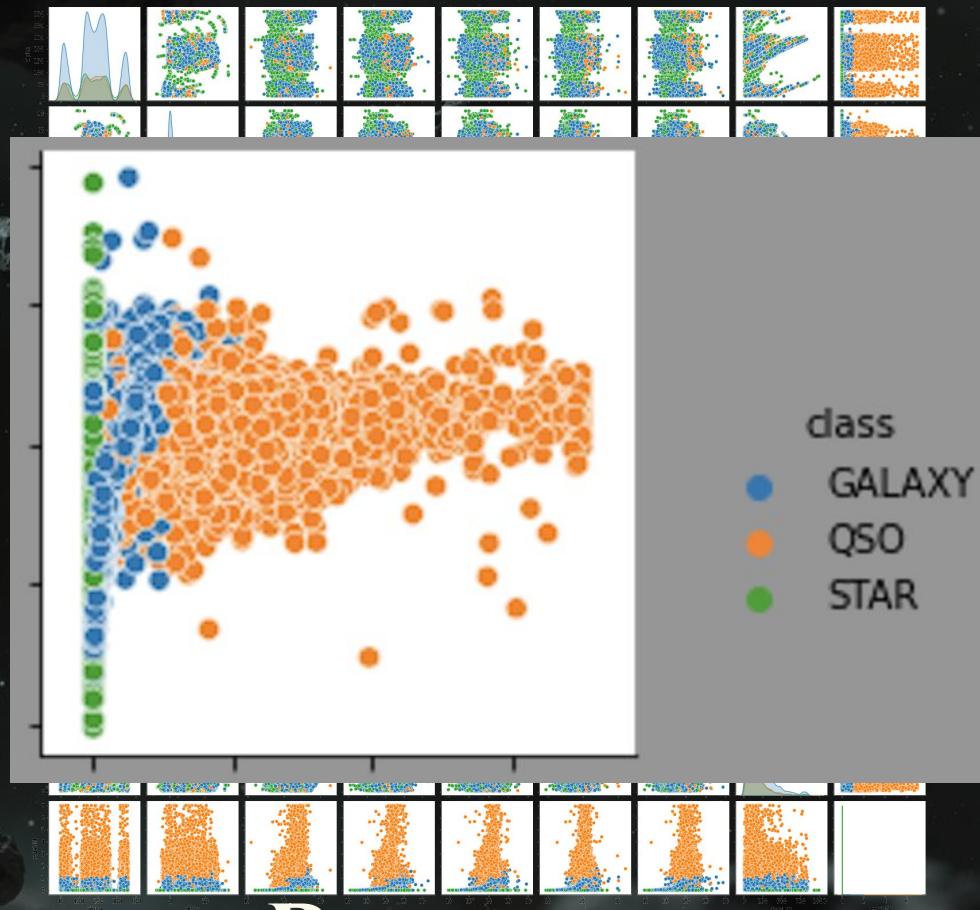
- **max\_features': 'sqrt'**
- **min\_samples\_leaf': 4**
- **n\_estimators': 944,**
- **'scaler': MinMaxScaler()**



# BASELINE PAIR PLOT

INTRODUCTION - METHOD - RESULTS - CONCLUSION - APPENDIX





# REDSHIFT

INTRODUCTION - METHOD - RESULTS - CONCLUSION - APPENDIX