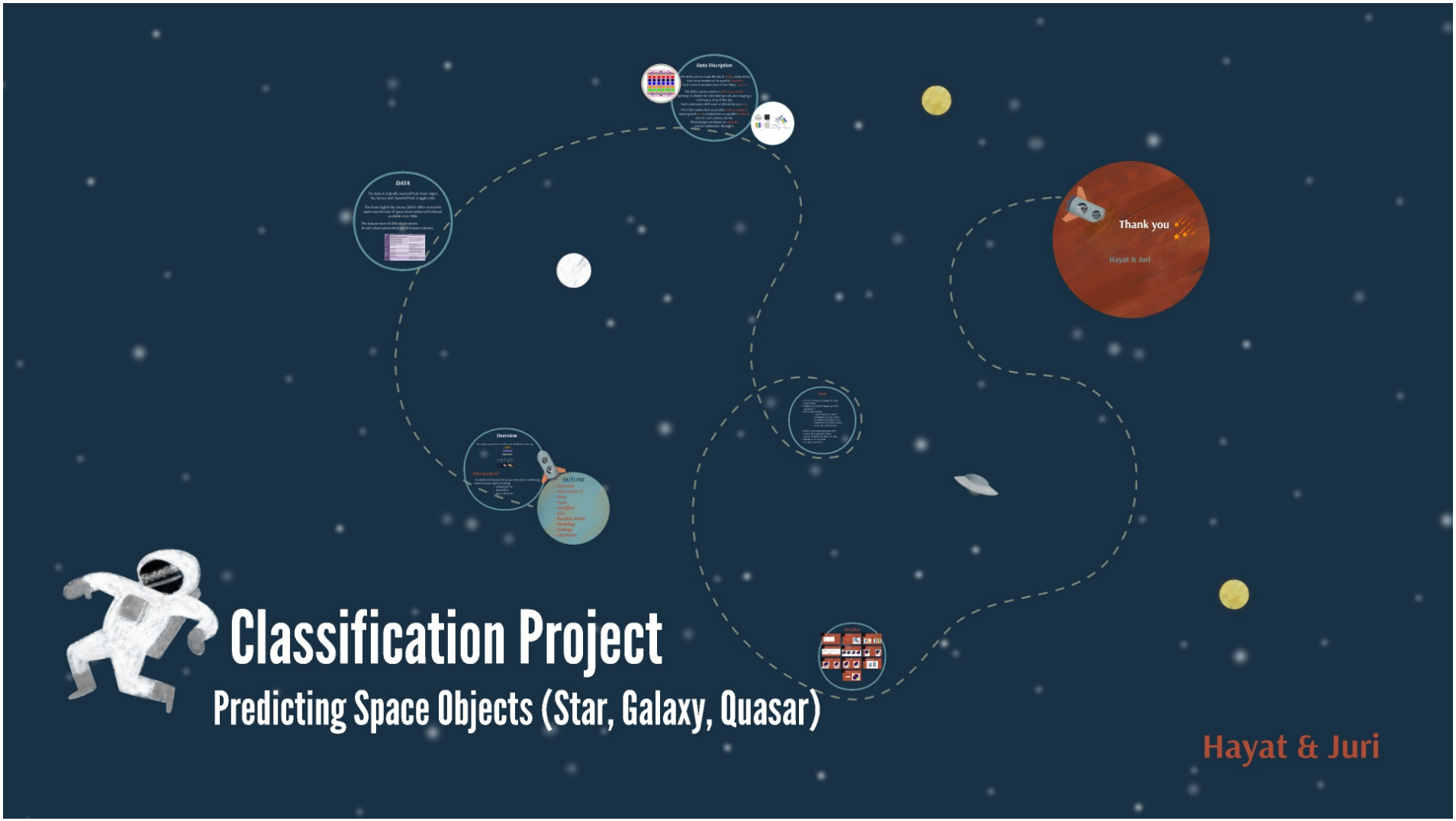

Data Description
DATA
Thank you
Hayat & Juri
Overview
OUTLINE
Classification Project
Predicting Space Objects (Star, Galaxy, Quasar)
Hayat & Juri


alaxies
uasars
Quasar
people interested in identifying
uding:
ysicists
ners
bservers
OUTLINE
Overview
Who needs it?
Data
Tools
Workflow
EDA
Baseline Model
Modeling
Findings
conclusion

Overview

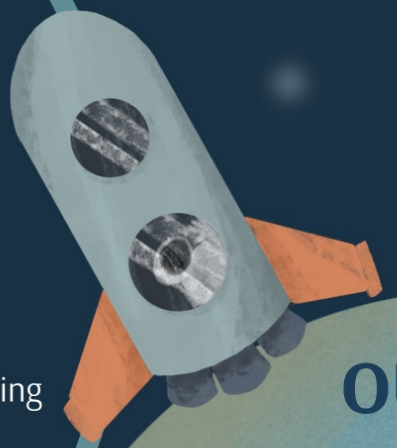
This project presents a classification model for predicting:

- Stars
- Galaxies
- Quasars

Who needs it?

The model will be useful for people interested in identifying observed space object including:

- astrophysicists
- researchers
- space observers



OUTLI

- Overview
- Who needs
- Data
- Tools
- Workflow

Star

A star is any massive self-luminous celestial body of gas that shines by radiation derived from its internal energy sources.

Galaxy

A galaxy is a sprawling systems of gas, dust, and billions of stars and their solar systems, all held together by gravity.

Quasar

an astronomical object of very high luminosity found in the centres of some galaxies and powered by gas spiraling at high velocity into an extremely large black hole.

Overview

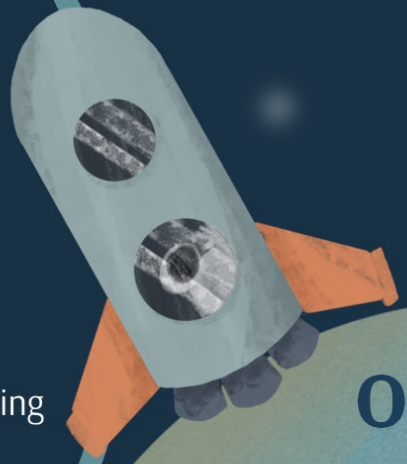
This project presents a classification model for predicting:

- **Stars**
- **Galaxies**
- **Quasars**

Who needs it?

The model will be useful for people interested in identifying observed space object including:

- astrophysicists
- researchers
- space observers



OUTL

- *Overview*
- *Who need*
- *Data*
- *Tools*
- *Workflow*

DATA

The data is originally sourced from Sloan Digital Sky Survey and imported from (Kaggle.com).

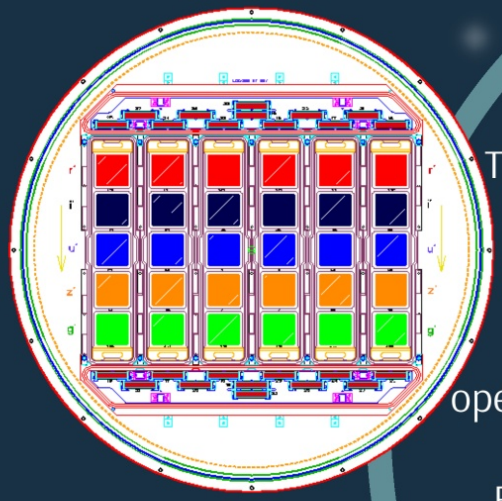
The Sloan Digital Sky Survey (SDSS) offers accessible open-sourced data of space observations with dataset available since 1988.

The dataset have 10,000 observations
In each observation there are 18 feature columns.

Features	Description	Remarks
objid	object Identifier	uniquely identify every object in data
Ra	J2000 Right Ascension (r-band), angular distance	
dec	J2000 Declination (r-band), angular distance	
u	better of DeV/Exp magnitude fit	Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope.
g	better of DeV/Exp magnitude fit	
r	better of DeV/Exp magnitude fit	
i	better of DeV/Exp magnitude fit	
z	better of DeV/Exp magnitude fit	
Run	Run Number	identifies the specific scan
Rereun	Rerun Number	Each rerun consists only in a change to the photometric pipeline, not to the underlying data
Camcol	Camera column	identifying the scanline within the run
Field	Field number	The field is an integer uniquely identifying a detection in the photo catalog
Specobjid	Object Identifier	uniquely identify every space object
Class	Space object class (galaxy, star, or quasar object)	
Redshift	Final Redshift	
Plate	plate number	
Mjd	MJD of observation	modified Julian date (gives the number of days since midnight on November 17, 1858)
Fiberid	Fiber ID	

Features	Description	Remarks
objid	object Identifier	uniquely identify every object in data
Ra	J2000 Right Ascension (r-band), angular distance	
dec	J2000 Declination (r-band), angular distance	
u	better of DeV/Exp magnitude fit	Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope.
g	better of DeV/Exp magnitude fit	
r	better of DeV/Exp magnitude fit	
i	better of DeV/Exp magnitude fit	
z	better of DeV/Exp magnitude fit	
Run	Run Number	identifies the specific scan
Rereun	Rerun Number	Each rerun consists only in a change to the photometric pipeline, not to the underlying data
Camcol	Camera column	identifying the scanline within the run
Field	Field number	The field is an integer uniquely identifying a detection in the photo catalog
Specobjid	Object Identifier	uniquely identify every space object
Class	Space object class (galaxy, star, or quasar object)	
Redshift	Final Redshift	
Plate	plate number	
Mjd	MJD of observation	modified Julian date (gives the number of days since midnight on November 17, 1858)
Fiberid	Fiber ID	

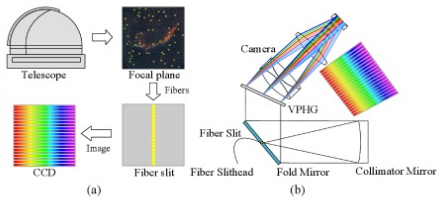
Data Discription



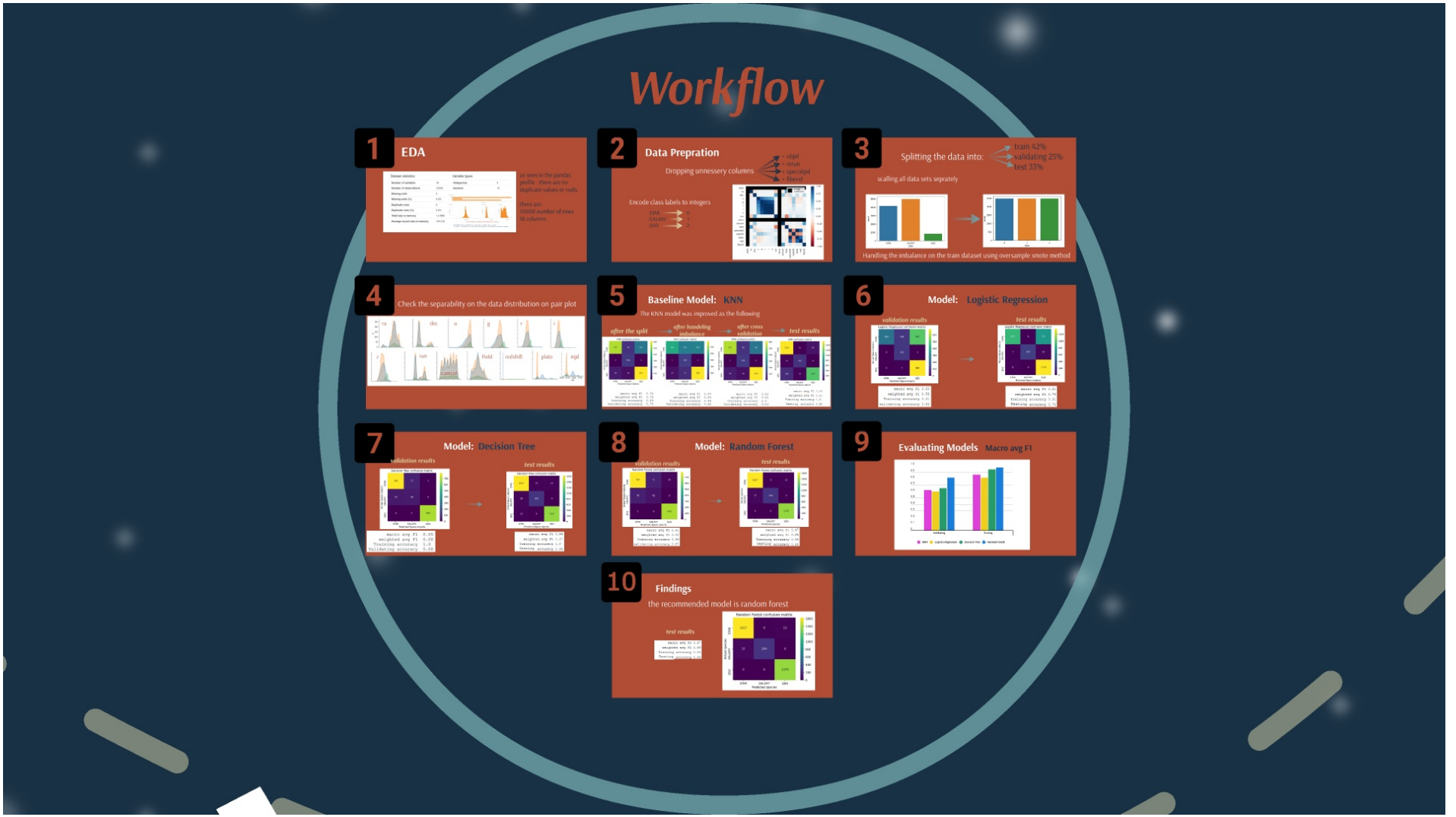
The SDSS camera scans the sky in **strips** along circles.
Each strip consists of six parallel **scanlines**.
Each scanline includes data in five filters, **u g r i z**.

The SDSS camera works in **drift scan mode** :
opening its shutter for extended periods and imaging a continuous strip of the sky.
Each continuous drift scan is referred to as a **run**.

The SDSS camera had six parallel **camera columns**,
meaning each **run** is divided into six parallel **scanlines**,
one for each camera column.
These images are known as **camcols**,
and are numbered 1 through 6.




Tools
Pandas and NumPy packages for data manipulation.
Seaborn & Matplotlib libraries for data visualization.
sklearn for modeling:
LogisticRegression model
KNeighborsClassifier model
DecisionTreeClassifier model
RandomForestClassifier model
train_test_split function
Imblearn for balancing the datasets.
Counter from collections library.
Jupyter notebook that hosts the code.
Tableau for visualization.
Prezi for presentation.


Workflow
1
EDA
as seen in the pandas profile - there are no duplicate values or nulls
there are 100000 number of rows 18 columns
2
Data Prepration
Dropping unnessery columns
objid
rerun
specobjid
fiberid
Encode class labels to integers
QSO
GALAXY
STAR
3
Splitting the data into:
train 42%
validating 25%
test 33%
scalling all data sets seprately
Handling the imbalance on the train dataset using oversample smote method
4
Check the separability on the data distribution on pair plot
5
Baseline Model: KNN
The KNN model was improved as the following
after the split
after handeling imbalance
after cross validation
test results
6
Model: Logistic Regression
validation results
test results
7
Model: Decision Tree
validation results
test results
8
Model: Random Forest
validation results
test results
9
Evaluating Models Macro avg F1
10
Findings
the recommended model is random forest
test results

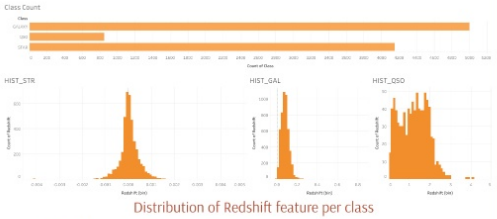
1 EDA

Dataset statistics

Number of variables	18
Number of observations	10000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.4 MiB
Average record size in memory	144.0 B

Variable types

Categorical	3
Numeric	15

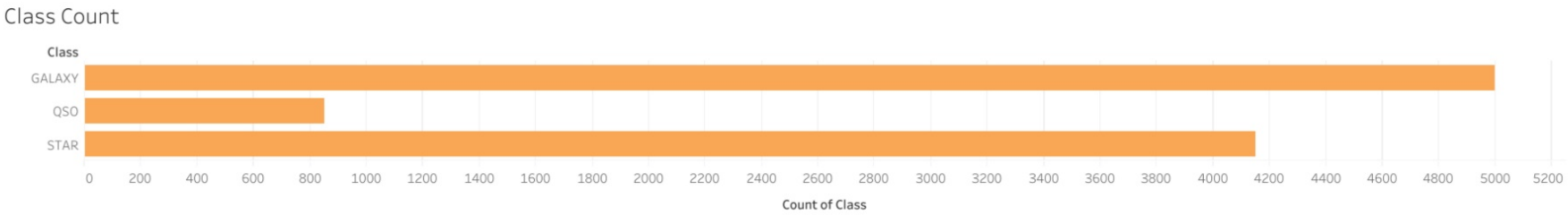


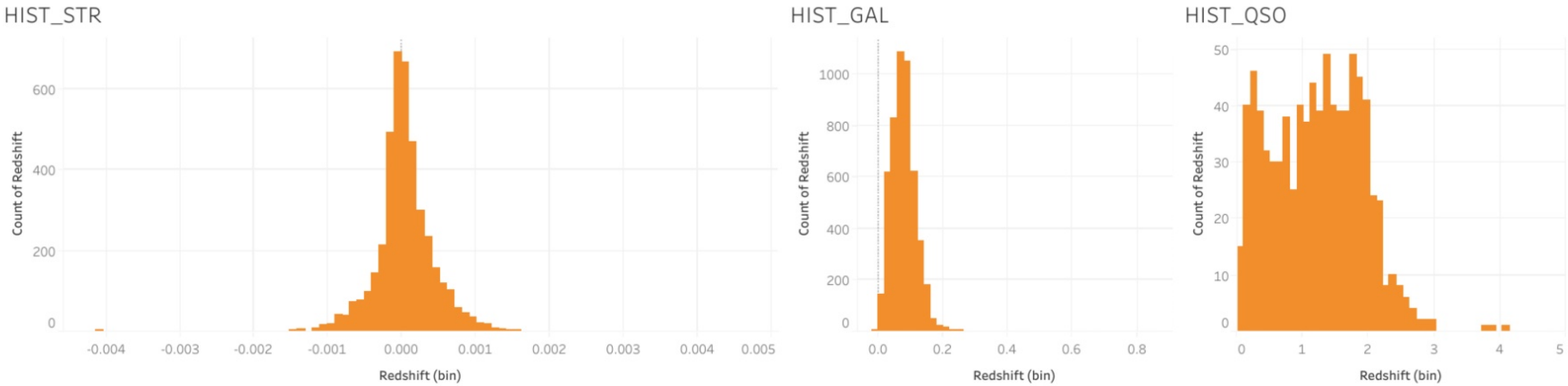
Distribution of Redshift feature per class

The redshift can be an estimate of the distance from the earth to a object in space
the plot tells us that most of the stars observed are somewhat closer to the earth than galaxies or quasars.

as seen in the pandas profile : there are no duplicate values or nulls.

there are:
10000 number of rows
18 columns





Distribution of Redshift feature per class

The redshift can be an estimate of the distance from the earth to a object in space
the plot tells us that most of the stars observed are somewhat closer to the earth than galaxies or quasars.

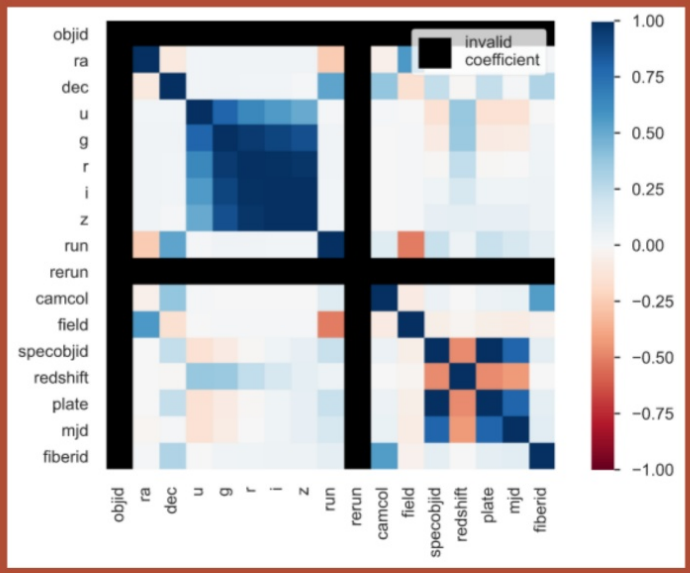
2 Data Prepration

Dropping unnessery columns

- objid
- rerun
- specobjid
- fiberid

Encode class labels to integers

STAR → 0
GALAXY → 1
QSO → 2

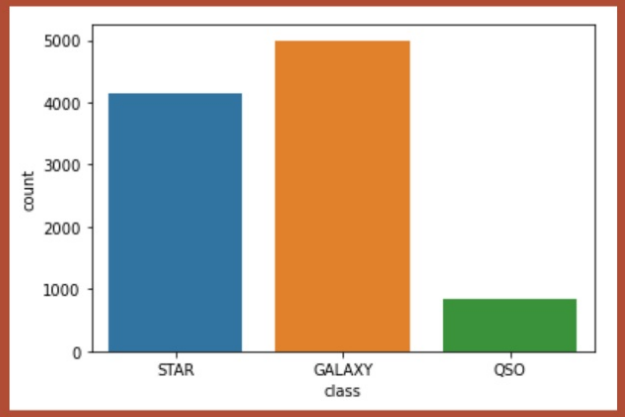


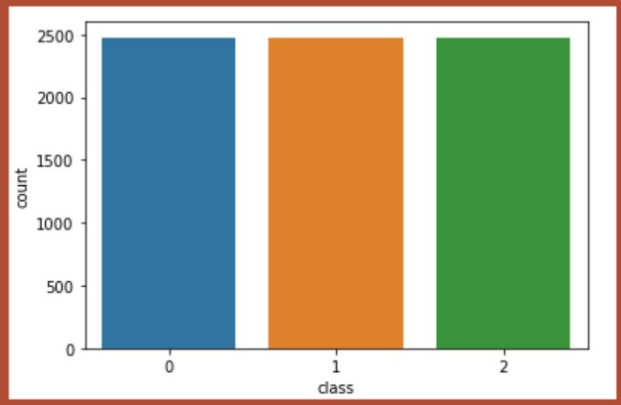
3

Splitting the data into:

- train 42%
- validating 25%
- test 33%

scalling all data sets seprately

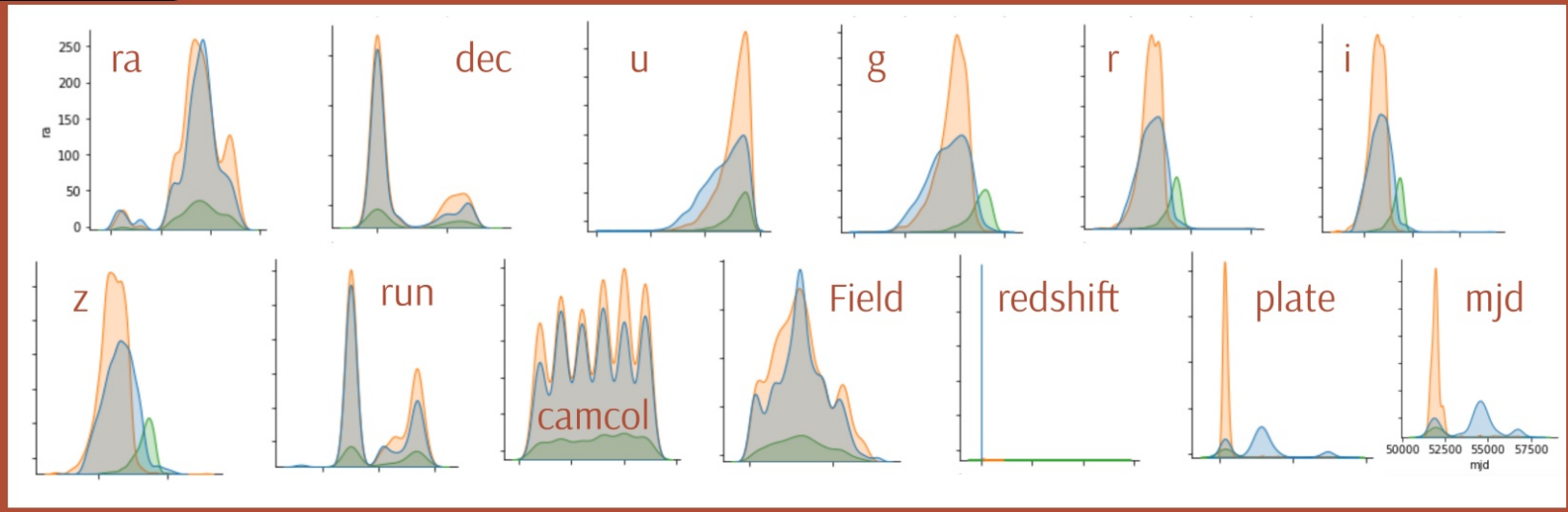




Handling the imbalance on the train dataset using oversample smote method

4

Check the separability on the data distribution on pair plot

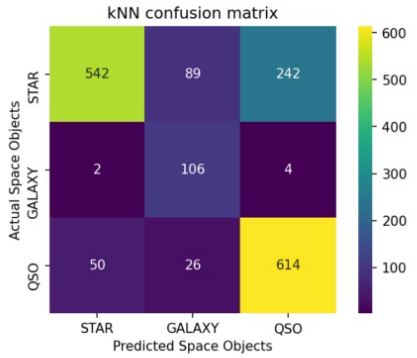


5

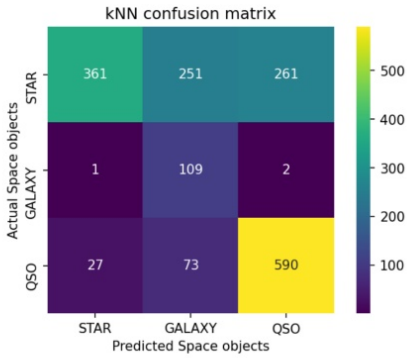
Baseline Model: KNN

The KNN model was improved as the following

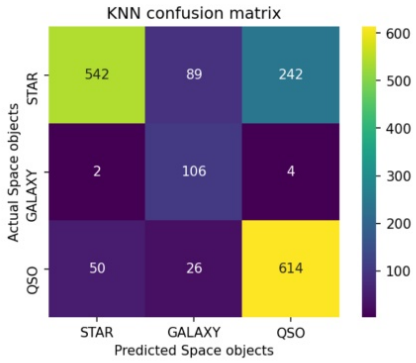
after the split → *after handeling imbalance* → *after cross validation* → *test results*



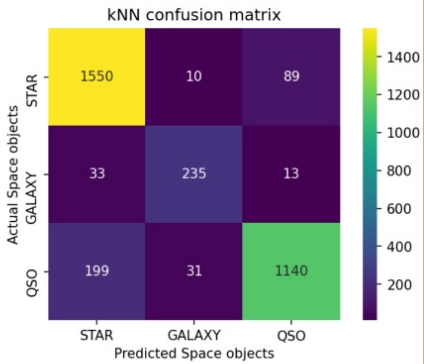

```
macro avg F1        0.72
weighted avg F1     0.75
Training accuracy   0.89
validating accuracy 0.75
```




```
macro avg F1        0.57
weighted avg F1     0.63
Training accuracy   0.94
Validating accuracy 0.62
```

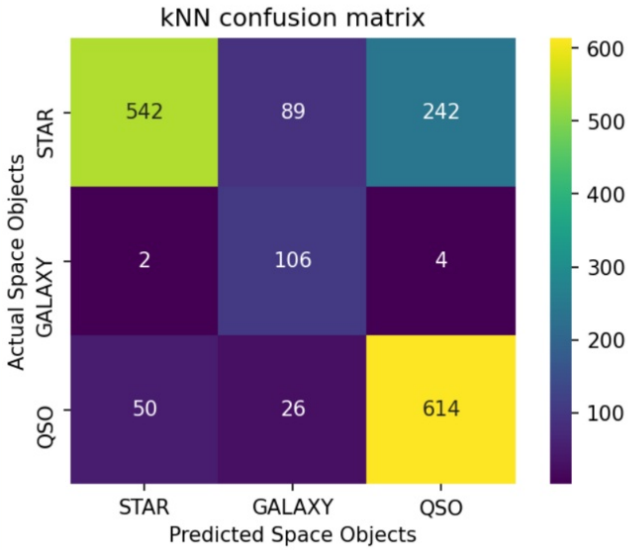



```
macro avg F1        0.62
weighted avg F1     0.62
Training accuracy   1.0
Validating accuracy 0.63
```

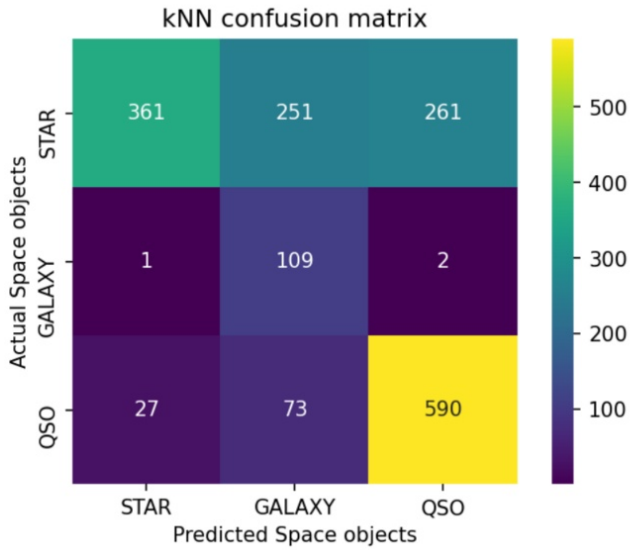



```
macro avg F1 0.86
weighted avg F1 0.88
Training accuracy 0.91
Testing accuracy 0.88
```

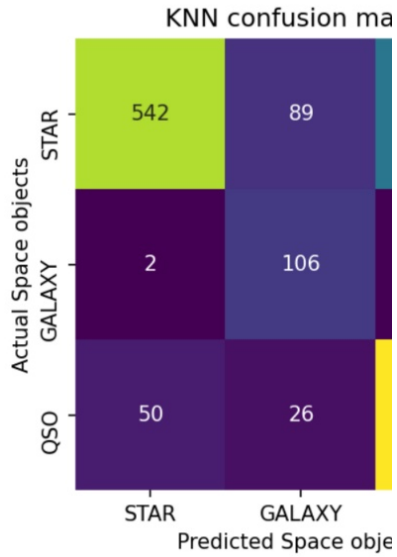
after the split → after handeling imbalance → after valida



macro avg F1	0.72
weighted avg F1	0.75
Training accuracy	0.89
validating accuracy	0.75

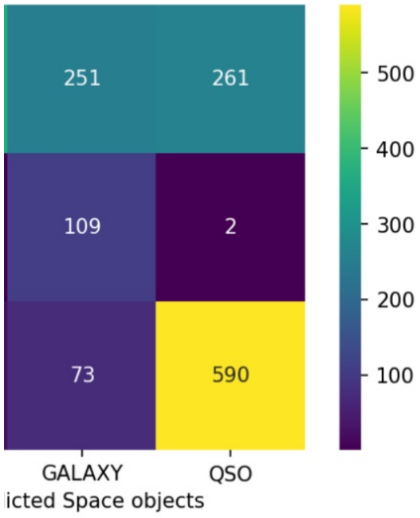


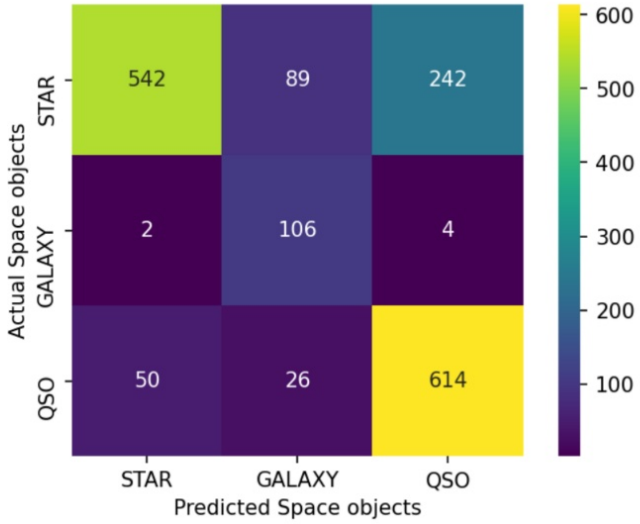
macro avg F1	0.57
weighted avg F1	0.63
Training accuracy	0.94
Validating accuracy	0.62

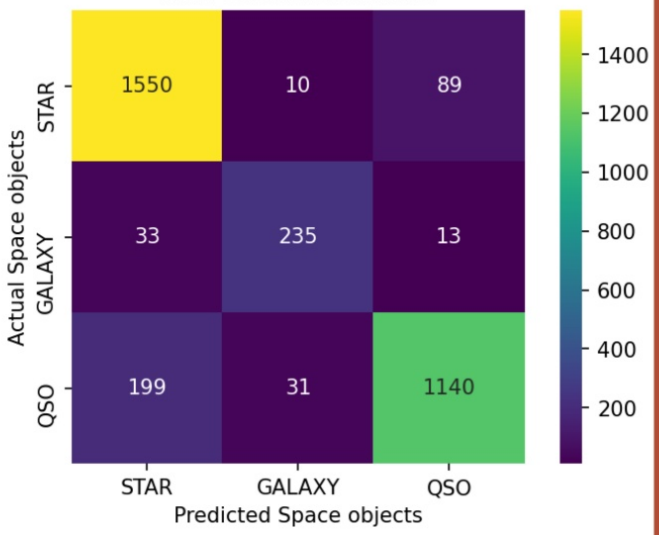


macro av	
weighted av	
Training accu	
Validating accu	

ter handeling imbalance → *after cross validation* → *test results*








```
 macro avg F1   0.57
.ghted avg F1   0.63
ıing accuracy   0.94
:ing accuracy   0.62
```

```
macro avg F1   0.62
weighted avg F1   0.62
Training accuracy   1.0
Validating accuracy   0.63
```

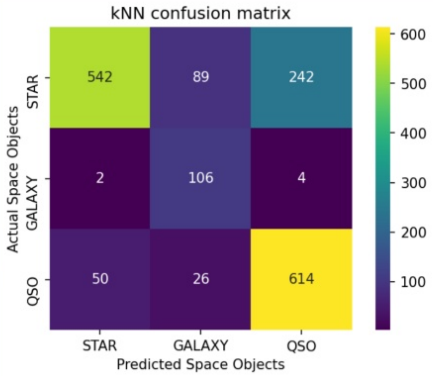
```
macro avg F1 0.86
weighted avg F1 0.88
Training accuracy 0.91
Testing accuracy 0.88
```

5

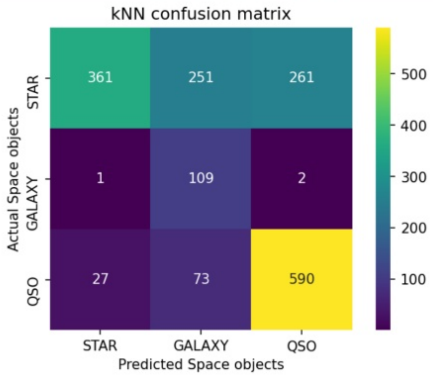
Baseline Model: KNN

The KNN model was improved as the following

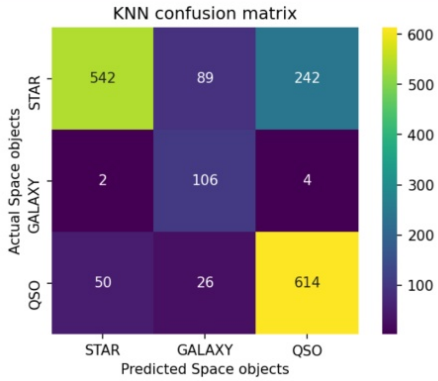
after the split → *after handeling imbalance* → *after cross validation* → *test results*



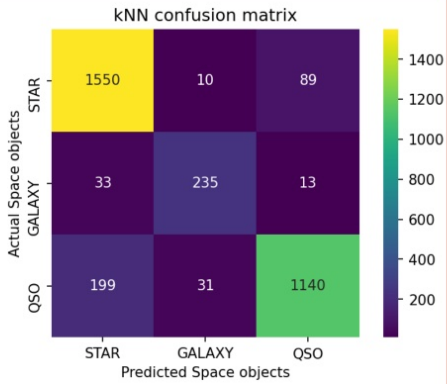

```
          macro avg F1  0.72
       weighted avg F1  0.75
    Training accuracy   0.89
validating accuracy     0.75
```




```
          macro avg F1  0.57
       weighted avg F1  0.63
    Training accuracy   0.94
   Validating accuracy  0.62
```




```
          macro avg F1  0.62
       weighted avg F1  0.62
    Training accuracy   1.0
   Validating accuracy  0.63
```

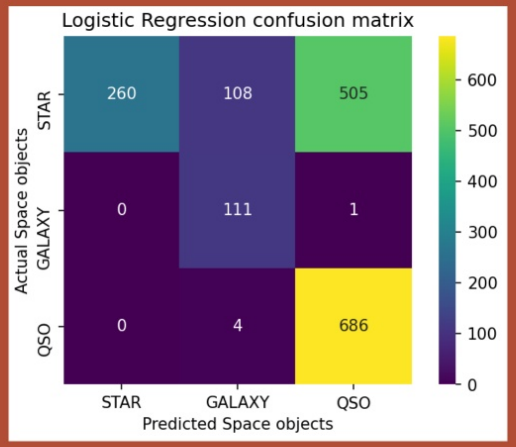



```
          macro avg F1  0.86
       weighted avg F1  0.88
    Training accuracy   0.91
    Testing accuracy    0.88
```

6

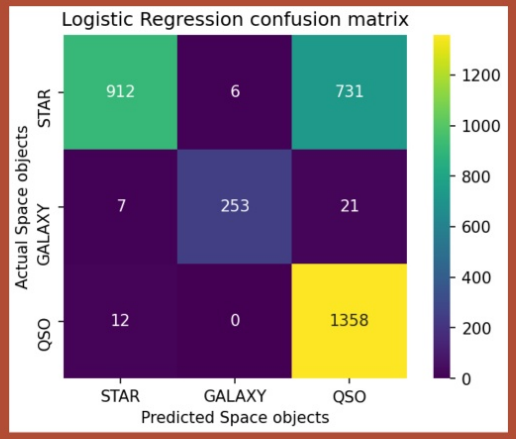
Model: Logistic Regression

validation results




```
macro avg F1 0.60
weighted avg F1 0.58
Training accuracy 0.91
validating accuracy 0.62
```

test results

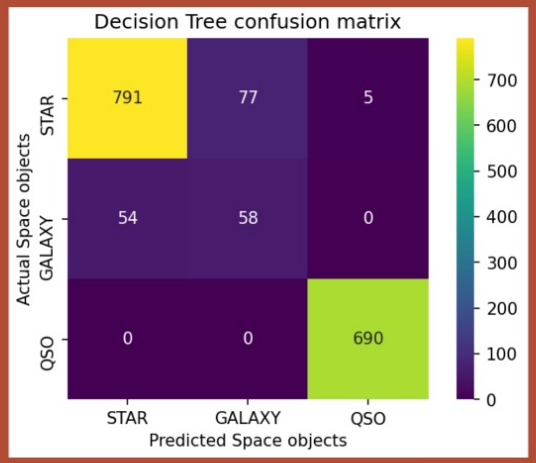



```
macro avg F1 0.81
weighted avg F1 0.76
Training accuracy 0.91
Testing accuracy 0.76
```

7

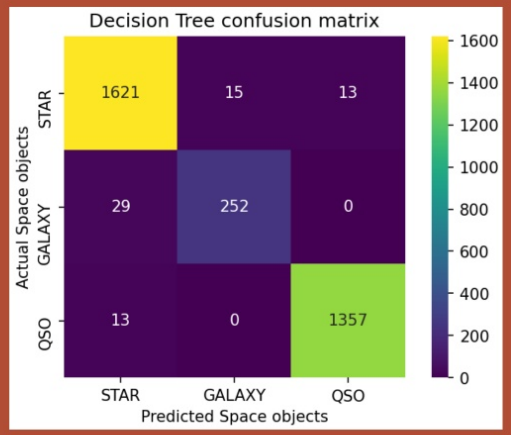
Model: Decision Tree

validation results



macro avg F1	0.65
weighted avg F1	0.88
Training accuracy	1.0
Validating accuracy	0.88

test results

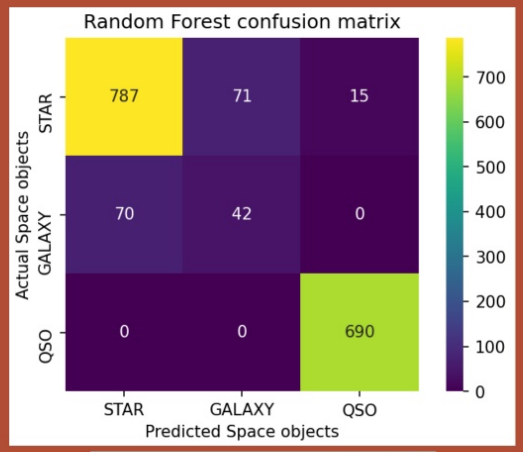


macro avg F1	0.94
weighted avg F1	0.97
Training accuracy	1.0
Testing accuracy	0.96

8

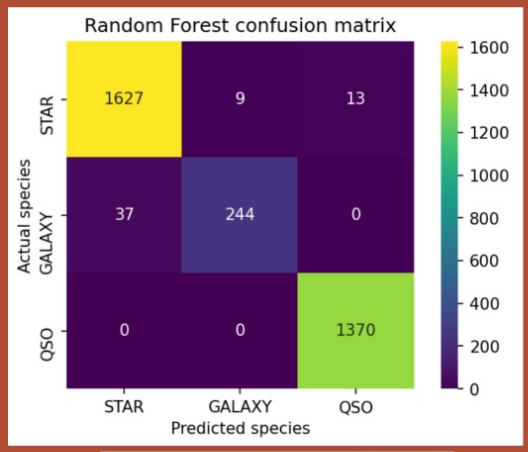
Model: Random Forest

validation results




```
macro avg F1 0.81
weighted avg F1 0.88
Training accuracy 0.99
validating accuracy 0.87
```

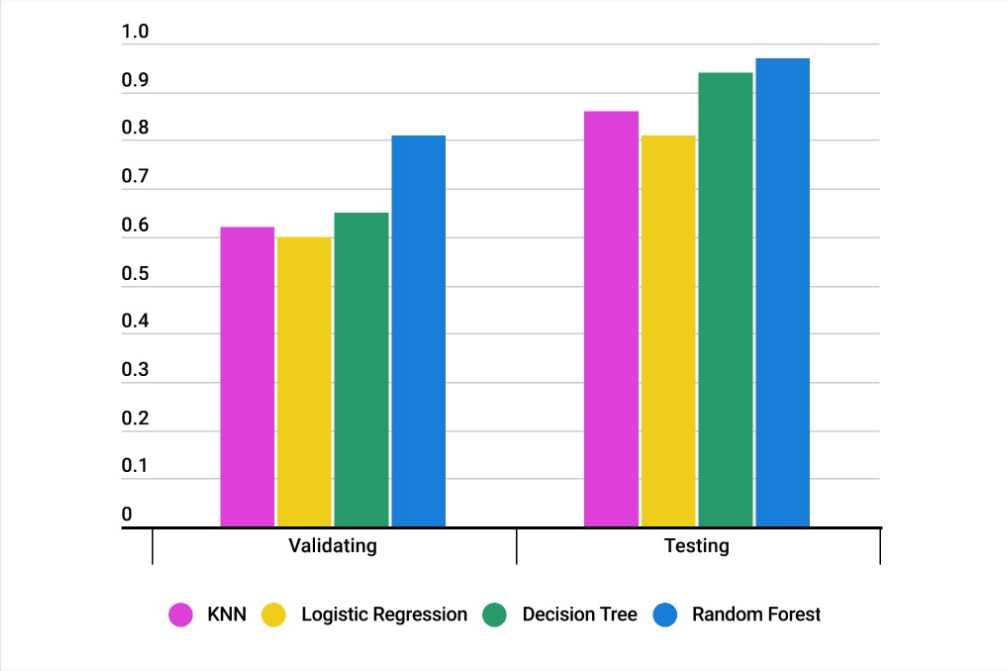
test results




```
macro avg F1 0.97
weighted avg F1 0.99
Training accuracy 0.99
Testing accuracy 0.98
```

9

Evaluating Models Macro avg F1



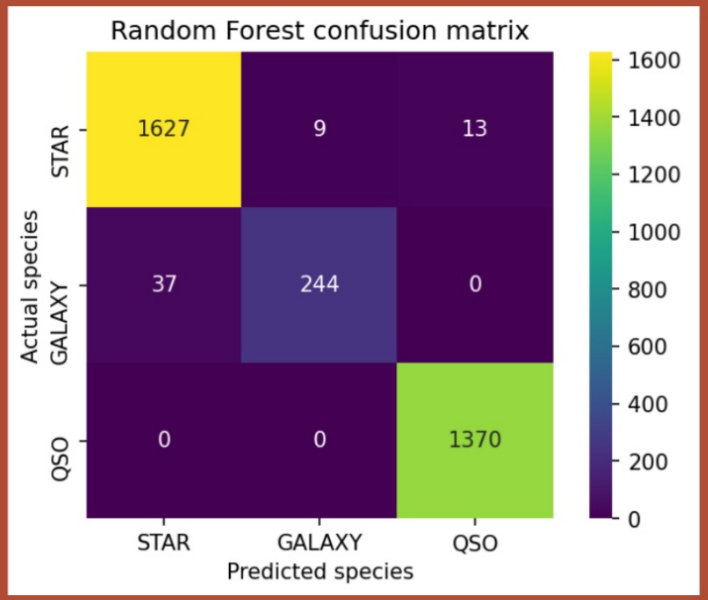
10

Findings

the recommended model is random forest

test results

macro avg F1	0.97
weighted avg F1	0.99
Training accuracy	0.99
Testing accuracy	0.98




Thank you
Hayat & Juri