Classification Project Report

Predicting Space Objects (Star, Galaxy, Quasar)

Hayat Aldhahri & Juri Alsayigh 10/28/21

أكاديمية سدايا
SDAIA Academy

Overview:

This project presents a classification model for predicting Stars, Galaxies and Quasars. The model will be useful for astrophysicists, researchers, space observers and people interested in identifying observed data from the space and classify them into either Starts, Galaxy or Quasars. 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

Design:

This project originates from the Data Science Bootcamp (T5) to build a classification Algorithms on space objects (Star, Galaxy, Quasar). The algorithms utilized were K-Nearest Neighbor, Logistic Regression, Random Forest and Decision Tree and compare results from these models.

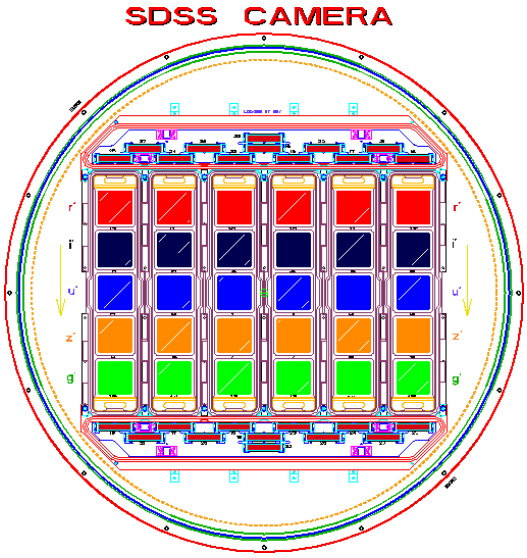
Data Description:

The dataset has 10,000 observations of space provided by the SDSS. In each observation there are 18 feature columns.

Table 1: Data Description

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	

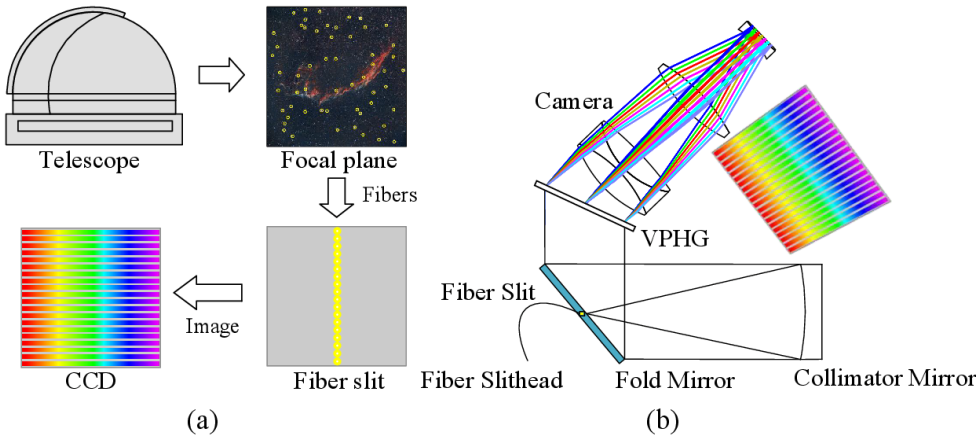
The SDSS camera scans the sky in strips along great circles. Each strip consists of six parallel *scanlines*. Each scanline includes data in all five filters, *ugriz*.



The SDSS camera worked in drift scan mode, opening its shutter for extended periods, and imaging a continuous strip of the sky. In the figure shown aside the sky drifts downwards. Each continuous drift scan is referred to as a run and there is a unique integer identifying the run.

The SDSS camera had six parallel camera columns, meaning that 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.

The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID. Each fiber is surrounded by a large sheath which prevents any pair of fibers from being placed too close on the same plate. When two targets are too close to each other the highest-priority target is observed. The other target may be observed if there is an overlapping plate covering this region.



Algorithm

Data Pre-Processing:

- Ensure data had no nulls/missing values.
- Ensure no duplicated values.
- Drop unnecessary columns such as (objid, rerun, specobjid, fibered).
- Encoding classes to integers 0 1 2 in order to be used in the model .
- Ensure data is balanced for the model.

أكاديمية سدايا
SDAIA Academy

Visualization:

ProfileReport was performed for EDA.

Dataset statistics

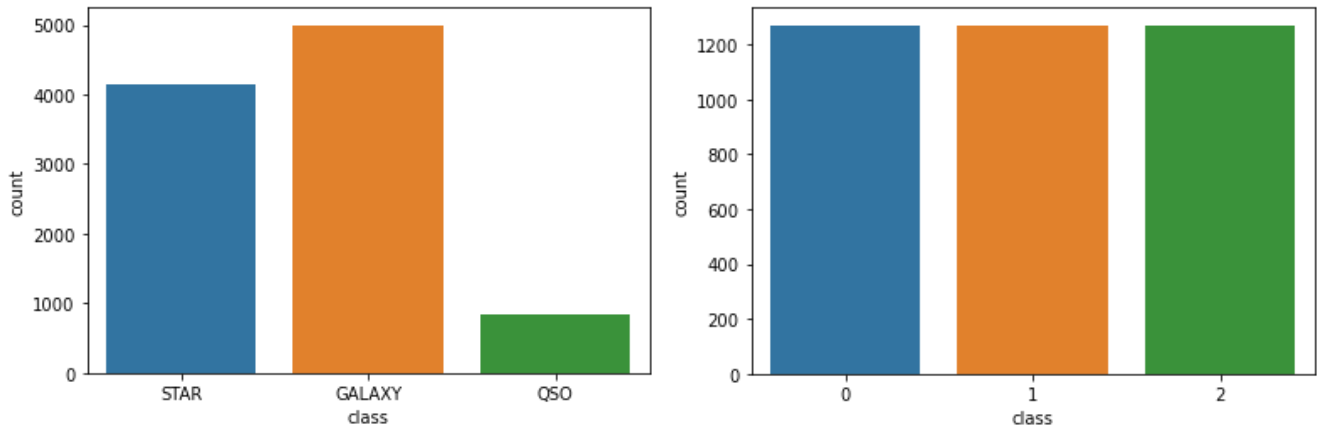
Number of variables	18
Number of observations	10000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.4 MiB
Average record size in memory	144.0 B

Variable types

Categorical	3
Numeric	15

Modeling and Evaluation:

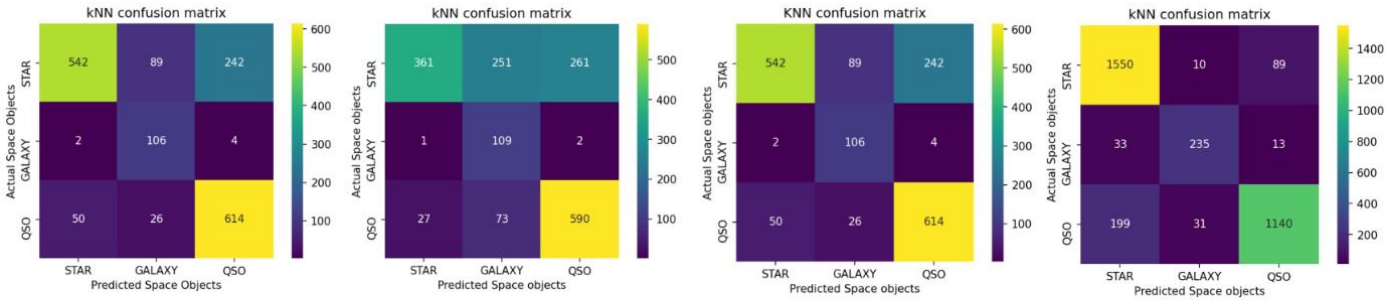
Our baseline model was KNN, and to ensure that the data is well balanced, smote over sample method was utilized for the class labels as shown below.



The data was split into:

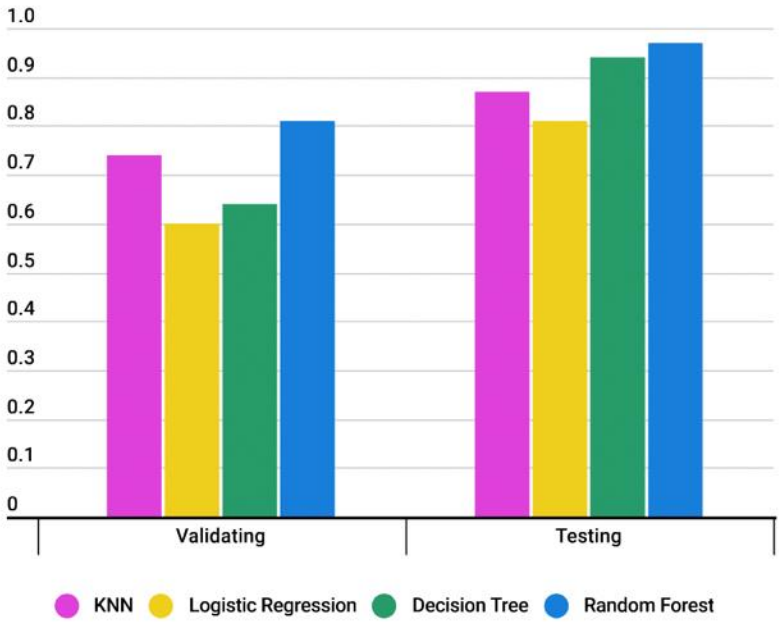
- 42% Train
- 25% Validating
- 33% Test

We started modeling the data using KNN. Initially, the training accuracy was approximately 89%. This was improved following the data split, handling imbalance and cross validation. The below graph illustrates the results.



macro avg F1	0.72	macro avg F1	0.57	macro avg F1	0.62	macro avg F1	0.86
weighted avg F1	0.75	weighted avg F1	0.63	weighted avg F1	0.62	weighted avg F1	0.88
Training accuracy	0.89	Training accuracy	0.94	Training accuracy	1.0	Training accuracy	0.91
validating accuracy	0.75	Validating accuracy	0.62	Validating accuracy	0.63	Testing accuracy	0.88

The below bar char surmises the F1 metric for the validation and testing of the four models. based on the results the random forest is the best classification model for our data.



Tools:

- Pandas and NumPy packages to data manipulation.
- Matplotlib library for data visualization.
- LogisticRegression model from sklearn.linear_model class for classification.
- train_test_split function in Sklearn model selection for data splitting.
- KNeighborsClassifier model from sklearn.neighbors
- DecisionTreeClassifier model from sklearn.tree
- RandomForestClassifier model from sklearn.ensemble for classification model.
- Jupyter notebook that hosts the code.
- Tableau for visualization.
- Prezi for presentation.

Communication:

Please refer to the presentation for more insights.
(https://prezi.com/view/Ue3ITBy52O9bY6iKTuDD/)
Code can be accessed in the following link to Github.
(https://github.com/Jurisayigh/Classification-Project/blob/main/Space%20Object%20Classification%20Project.ipynb)
(https://github.com/hayataldhahri/Classfication-Project)

Also Tableau dashboard can be accessed through the following public link (https://public.tableau.com/app/profile/hayat4538/viz/SDSSData_StarClassification/Dashboard1)