**Course**: DATASCI207 – Applied Machine Learning, Section 7

**Instructor**: Ishaani Priyadarshini

**Assignment**: Comprehensive Project

**Title:** Unlocking the Cosmos: Machine Learning Insights into Astronomical Object Classification and Galaxy Typing from SDSS and Galaxy Zoo 2

**Authors**: Ahmad Allaou, Daelyn Bergsman, and Spencer Morris

**Date:** 14 December 2023

**GitHub Repository**: https://github.com/jspencermorris/datasci207\_project

**Contents**:

* Motivation
* Objectives
* Classification Schema
  + *Superclass*
    - *Galaxies and Hubble Subclassifications*
    - *Quasars*
    - *Stars and HZ Subclassifications*
* SDSS Project
* Galaxy Zoo Project
* Datasets
* Evaluation Methodology
* Tabular Preprocessing
  + *Filtering*
  + *Label Engineering*
  + *Feature Engineering*
  + *Splitting*
  + *Balancing Approaches*
  + *Transformations*
  + *Summary*
* Image Preprocessing
* Model Details
* Classification Results
* Discussion
  + *Summary of conclusions*
  + *Limitations*
  + *Future Work*

**Color Contexts for Editing Document**

Yellow = AA/DB – image write-ups

Gray = JSM – tabular write-ups

Red = Combined – final write-ups

Cyan = references – re-sequence

Brown = supporting figures – re-sequence

**Motivation**

Advances in data analysis and astronomical research have been tightly intertwined since the advent of the method of least squares regression and its early application to observations of planetary motion [A]. In modern times, big-data astronomical projects have greatly expanded our ability to explore the cosmic landscape. However, the sheer volume and complexity of data generated, particularly by observatories conducting wide-field surveys, necessitate advanced data processing pipelines, and highlight the crucial role of automated systems in data reduction, pattern recognition, and classification. The next-generation Vera C. Rubin Observatory, for example, is anticipated to publish 5 petabytes per year, much of which must be labeled in real-time using a system of data processing pipelines [B].

This paper provides a demonstration of celestial object classification using a suite of machine learning models as applied to data of various shapes generated from the Sloan Digital Sky Survey (SDSS) during its initial phase. By exploring both supervised and unsupervised approaches, we hope to exhibit methods available to machine learning practitioners as well as contribute to the broader discussion of big-data astronomy.

**Objectives:**

We aim to evaluate a subset of SDSS survey objects and develop models to classify them as either a galaxy, quasar, or star. For those survey objects classified as stars, we will further subclassify them based on a stellar classification system. We will model the predicted temperature of such stars using remaining metric features. For those survey objects classified as galaxies, we will attempt to further subclassify them based on morphological labels of their corresponding images. In particular, we aim to classify galaxy images according to morphological labels ascribed by the Galaxy Zoo Projects. In summary, the work herein is primarily concerned with 3 classification tasks: superclass, star subclass, and galaxy subclass.

Given the large size of the datasets and the data engineering and computational challenges with developing our first machine learning models, we aim to develop models that are reasonably performant. We hope to achieve a high accuracy (at or above 90%) for our first two objectives, but suspect the image classification task will be considerably more challenging, given the diffuse shapes of galaxies and limited resolution of available images.

**Classification Schema**

Within the observable universe, matter accumulates as astronomical structures across a wide range of distance scales, from as large as ~1024 meters (such as cosmic filaments) down to as small as ~10 meters (such as asteroids). In this paper we restrict ourselves to three **superclasses**: galaxies, quasars, and stars.

**Galaxies** are vast agglomerations of matter with masses ~1012 solar masses and lengths ~1020 meters. Other than the band of stars that comprise the disk of our own Milky Way, only a few other galaxies are visible to the naked eye – the most prominent being the nearby giant Andromeda. These objects are the primary features visible in deep-field astronomical images, and the observable universe is estimated to contain ~1012 of them. The prevailing cosmological model, Lambda-CDM, predicts formation of primordial galaxies from collapsing dust in the early universe when it was much denser, although this is an area of active research and observations are not fully explained.

Regardless of the details of their formation, galaxies seem to exhibit a limited set of morphological features, which was first described using a ‘tuning-fork’ schematic colloquially known as the Hubble Sequence [1]. The scheme contains 5 main classifications: irregular (which are devoid of a well-organized shape), elliptical (which are spheroidal), lenticular (which are disk-shaped and often have a central spheroidal bulge but generally otherwise lack substructure), spiral (which are disk-shaped, often have a central bulge, and consist of a small number of dense spiral arms), and spiral-barred (which are similar to regular spirals with the exception that the central bulge is extended into a prominent bar and typically have two primary arms). Various intermediate and extended classifications exist beyond the simple Hubble Scheme that are descriptive of elliptical eccentricity and the number and compactness of spiral arms. In this study, we restrict ourselves to a simple version of the Hubble Sequence consisting of just 15 subclasses.

**Quasars** are a further subclassification of galaxies and are examples of active galactic nucleii (AGN). These objects are observed to be significantly more luminous and distant than normal galaxies. As there are relatively few of them and they hold clues to the composition of early galaxies, they are of special interest to cosmologists.

Other than the planets and the moon, the primary features of the night sky are nearby **stars**. Of the ~1011 stars that collectively comprise the Milky Way galaxy, only a few thousand are visible to the naked eye. Stars represent the spheroidal accumulation of cosmic dust of ~10 solar masses into diameters of ~1010 meters. As they possess the densities necessary for nucleosynthesis, they are the primary source of electromagnetic radiation (and thus originate most of the universe’s visible light).

The lifecycle of most stars is currently best understood as an interplay of mass, density, and age, being visualized as a ‘main sequence’ along the primary dimensions of luminosity and color index and exemplified by the Hertzprung-Russell Diagram [25]. A star’s color index, therefore, is its foundational classification, and has been formalized using the Morgan-Keenan (MK) system [2]. In the MK system, a star’s spectrum is used to infer the presence of chemical species and surface temperature, which is then mapped to one of 10 several letters (O being very hot and T being very cold) [3]. Furthermore, the MK system further subclassifies stars based on their luminosity, which is mapped to one of several Roman numerals. In this study, we restrict ourselves to a simple version of the MK system consisting of 9 subclasses.

**SDSS Project**

The Sloan Digital Sky Survey (SDSS) is a pioneering multi-spectral imaging survey of ~26% of the sky with coverage primarily in the Northern Hemisphere [C]. Initially commissioned to develop a uniform, well-calibrated map to study the large-scale structure of the universe, SDSS data acquisition began in 2000, and has since gone through numerous extension phases, characterizing the universe in an increasingly comprehensive way [D]. It remains operational to this day and has contributed to a wide range of basic science including the stellar composition of the Milky Way, supernovae distributions, expansion properties of the universe, exoplanet discoveries by radial velocity, galaxy structure and evolution, and supermassive black holes. Furthermore, as one of the first big data astronomical projects, it has contributed to many advances in data storage, reduction, and accessing technologies.

In its original ‘Legacy’ phase, SDSS consisted of a dedicated 2.5 meter primary telescope, a primary imaging camera containing 30 photometric CCD’s with 5 color filters (spanning ultraviolet to deep infrared) [17], and two double spectrographs (with a combined wavelength range of 3600 to 10400 Å) paired to CCD’s [E]. Even at inception, therefore, SDSS provided a rich variety of data structures, including 0-d scalars (photometric brightness under a given color filter), 1-d vectors (detailed spectrum), and 3-d tensors (multispectral images).

Among other achievements, the Legacy SDSS data pipeline automatically classified objects by estimating point spread functions (PSF), which is a measure of the diffusiveness of an imaged object [F].

**Galaxy Zoo Project**

The early SDSS Legacy study ultimately characterized ~million galaxies via photometry, spectroscopy, and imaging, enabling galaxy studies of unprecedented scope. The image catalog, in particular, afforded a unique opportunity to interrogate the distribution of galaxy morphologies.

This deluge of data inspired the Galaxy Zoo Project (GZ1), which in turn spawned the Zooniverse initiative whose aim is to support labeling expansive scientific datasets by crowdsourcing labeling tasks to citizen scientists. GZ1 permitted 4 labels (elliptical, spiral, star/don’t know, and merger), and ultimately led to ~107 individual classifications by ~105 participants of ~250 thousand galaxies.

Following GZ1, a Galaxy Zoo 2 Project (GZ2) was launched that utilized a decision-tree framework and allowed for assignment of 37 binary labels (such as ordinal eccentricity scores for ellipticals and measures of the number of spiral arms for spirals) to ~300 thousand galaxy images [21]. Studies subsequent to the original paper have computed quality scores based on robust statistics, such as a ‘clean’ flag associated with images for which at least 10 votes were cast and for which >80% of voters (themselves weighted by their trustworthiness) agreed on a single label. Accompanying photometric and spectroscopic allow for close inspection of the largest (eg. Petrosian Radius > 17 arcsec) and highest-magnitude (eg. r-band Petrosian Magnitude < 17 maggies) objects, as well as advanced de-biasing techniques [H].

**Datasets**

In order to ensure data provenance, all data were retrieved from their original sources. The Galaxy Zoo 2 data (including the images, assignment statistics, and final labels) were obtained via the [project’s website](https://data.galaxyzoo.org/), and SDSS photometric/spectroscopic survey data were accessed via the Catalog Archive Server Jobs System (CasJobs).

The Galaxy Zoo 2 data came packaged as a zip file containing over 240k thousand images (as catalogued in the Hart paper), along with two tables [G]. The first table included a simple mapping between SDSS DR7 ID and image filename. The second table included 37 binary labels corresponding to the 37 possible answers from the GZ2 decision tree, 37 associated probabilities based on unweighted user responses, and 37 debiased user responses, along with a single categorical label corresponding to the best consensus classification. Because the latter categorical classification took on 818 possible values, a label-engineering process was performed to compress those values into the 14 labels traditionally found in the Hubble Sequence diagram. Although this process resulted in a greatly-simplified classification task, it resulted in loss of some additional labels, such as the ‘boxiness’ of the galactic bulge and the number of spiral arms.

One million records corresponding to suspected galaxies from SDSS Digital Release 16 (DR16) were obtained using a custom SQL query via CasJobs. Multiple tables were joined to return the following raw and processed columns populated from several sensor data pipelines: 5 identifiers, 12 Galaxy Zoo 1 labels, 6 classes, 20 numerical features, and 2 quality fields. To ensure the galaxy objects under evaluation were consistent with the Galaxy Zoo paper (such that they were sufficiently large for high-quality sensor and image data), results were limited to objects with Petrosian Radii (calculated for the red band) greater than 17.0 arcseconds. A second custom SQL query was written to retrieve one million records from SDS DR16 corresponding to objects labeled as quasars or stars, with a similar set of returned columns (sans the galaxy-specific fields).

**Evaluation Methodology**

Prototyping was primarily conducted using Jupyter Notebooks running Python 3.10 and various packages managed in a virtual conda environment. Key packages included pandas, numpy, matplotlib, seaborn, sklearn, tensorflow, and keras.

The GitHub repository contains most of the key files necessary for replication of this study. The full set of environment dependencies are documented in requirements.txt, which is available in the main directory. The original datasets were stored locally or in the Cloud during analysis, but a compressed file containing the consolidated and cleaned tabular data is available in /data/processed. Those models judged to be best, corresponding to each classification task and modeling paradigm, were pickled and saved in /models. The /notebooks folder contains a copy of all notebooks, including those used for exploratory data analysis, preprocessing, and modeling. The /references folder contains key literature consulted in development of this project. The /src/data folder contains the SDSS SQL queries. Finally, all finalized work is contained in /reports.

Modeling of tabular data was conducted locally on a machine with an Intel Core i7-7660U CPU, Iris Plus 640 integrated GPU, and 16 GB of DDR3 RAM. Run-times of tabular model fits were recorded for this configuration. Modeling of imaging data was conducted using a free Google Colab instance (with Fall 2023 runtime environment) in order to take advantage of parallel processing for improved performance.

All model hyperparameter tuning was performed strictly using the train and validation subsets, and only after the most performant models were selected were metrics collected for the train set.

Metrics for model evaluation (baselines for accuracy, definitions/interpretations/relevance of key metrics, …

**Tabular Preprocessing**

All tables were joined together using the DR7 and DR8 object ID’s prior to filtering.

Because the data were obtained from multiple sensors and data acquisition pipelines, objects had multiple label variables determined by different classification procedures. A derived label for ‘superclass’ was added and populated based on two strict criteria: labels were not contradictory and the object was flagged by SDSS as ‘clean’ (i.e. no documented artifacts or errors during data acquisition). After removal of duplicates, the resulting set of data with these high-fidelity labels included ~1.7 million objects.

As seen in the following image of their spatial distribution, objects were primarily located in the Northern Hemisphere with right ascensions between 110 and 260 degrees [4].

Objects were dropped that included any of the following: non-physical sensor readings or computed features (such as a negative Petrosian Radius); quality scores (which is an SDSS computed feature that estimates the overall data integrity at the time of data processing) less than 0.8; contradictory label or otherwise flagged as ‘unclassified.’ The resulting cleaned dataset contained ~1.09 million objects with a ‘superclass’ label, ~220 thousand galaxies with a Hubble Sequence label, and ~120 thousand stars with a MK label. We considered dropping outliers based on Tukey or Mahanabolis measures but opted to include such data to retain variation.

Preliminary inspection of the resulting consolidated tabular dataset revealed 177 unique labels for stars (using the expanded MK system and as determined from the SDSS dataset) and 815 unique labels for galaxies (using the expanded Hubble Sequence and as determined from the Galaxy Zoo 2 dataset). Furthermore, the class sets were highly unbalanced: the largest galaxy class, for example, included over 20 thousand examples, in contrast to the smallest galaxy class which contained only 1 example.

A label engineering exercise was performed in order to improve the uniformity of label counts and make the classification tasks in this paper tractable, given currently-available resources. Galaxy classifications were mapped to a reduced set of 15 classes in line with the simplified Hubble Sequence, and stellar classifications were mapped to a reduced set of 9 classes in line with the simplified MK system [5,6].

The final label distributions were extremely unbalanced, as seen in the following summaries [7,8,9]. While not unexpected, given that such distributions aren’t uniform in nature, nor will any instrumentation be equally sensitive to all astronomical objects, the unbalanced nature of the dataset does present a challenge for machine learning algorithms, which are most successful when training sets include equally spaced label (in the case of regression) or equally represented labels (in the case of classification). The unbalanced nature of the dataset is discussed and addressed later.

18 of the original variables were explored for their potential usefulness as model features. Their distributions and covariances were visualized in aggregate and based on the 3 classification tasks. Ultimately, five features were selected for retention for all classification tasks.

Four photometric features were selected: the sum of all photometric measures of all 5 filters, the value of u-r (a measure of how ultraviolet the object appears, relative to red), the value of g-i (a measure of how green an object appears, relative to near-infrared), and the value of r-z (a measure of how red an object appears, relative to deep-infrared). All of these were derived features and allowed for an overall reduction in the size of the feature set. Their inclusion is furthermore reasonable on physical grounds, since it is well-known that the color of astronomical objects is often suitable alone for classification. As is evident in the histograms of these features, quasars are generally more luminous and bluer, whereas galaxies tend to be fainter and bluer [10,11,12,13].

Finally, the redshift value, as determined from each object’s spectrum, was selected. Redshift, when log transformed, was determined to be exceptionally well-resolved across superclass labels [14], and likely alone would do a fine job of classifying objects by superclass.

A variance-covariance heatmap of the selected features is shown for all objects, as is a pairplot with indication of superclass [15, 16].

In addition to the above 5 features, the 3 Elodie columns (Color, Temperature, and Metallicity) were included as additional variables specifically for modeling stellar objects. As such, two feature sets were explored (relevant to either superclass or subclass tasks), and all selected features were metric.

Following initial preprocessing, random subsets were defined with allocations of 60% to a train set, 20% to a validation set, and 20% to a test set.

In order to address extreme label imbalances and improve object classification fairness, a common hyperparameter was used to immediately rebalance the train set by either undersampling or oversampling.

The undersampling approach involved identifying the number of objects of the rarest class, as defined by a minimum set size, then randomly selecting the same number from all other classes. We fixed all evaluations at a minimum set size of 1000 example objects, which reduced the number of evaluated star labels to 5 and the number of evaluated galaxy labels to 10. This transformed train set included ~1.7 thousand examples for each stellar subclass and ~1.6 thousand examples for each galaxy subclass.

The oversampling approach employed the synthetic minority over-sampling technique (SMOTE). SMOTE is a data augmentation technique that generates a transformed initial train set by oversampling minority classes. The up-sampled train set included ~28 thousand examples for each star subclass and ~26 thousand examples for each galaxy subclass.

Several other transformations approaches were explored, though we comment on only those most noteworthy.

First, a transformation set X0 was generated using z-standardization of all five of the engineered features. Other than the clustering model, we used this set for all other subsequent models.

Next, a transformation set X3 was generated with, in addition to z-standardization, applied manual feature weights based on their origin from the instrumental data acquisition: in particular, the spectroscopic feature was given a weight of 0.5, and the remaining four photometric features were each weighted according to the remainder, at 0.25 each.

Finally, data reduction transformations of the two feature sets (i.e. the 5 features available to superclass and galaxy sublcass, and the 8 features available to the stellar subclass), X7, were generated using **principal component** models of the metric features. Over 90% of the variation within the 5 common features is explained by the first two principal components [X], and the redshift and photometric sum dominate those variable loadings [X]. Projection of the objects along the first 3 components shows good separation by superclass [X]. For the stellar feature set, ~85% of the underlying variation was explained using just two principal components [Y], where the temperature, photometric sum, metallicity, and redshift had the highest loadings [Y]. A scatter plot of the first two principal components allows for visualization of how this unsupervised learning model can be useful for stellar classification [Y].

The diagram of the preprocessing pipeline summarizes the overall approach used for the tabular dataset [22].

**Image Preprocessing**

……………………….

**Model Details**

**Simple linear regression** is…Results are…

**K-means** is an unsupervised learning model of hypershperical clusters of objects based on minimization of a function of the mean of majority-voted pairwise Euclidean distances, is prototypical of clustering, is computationally efficient, and requires an a priori assumption of the number of clusters within the dataset. Results are… K-means of X0, X3, and X7…

**The Gaussian mixture model is**… Results are…

**The naïve Bayes model is**… Results are…

**The K-nearest neighbors** model exemplifies supervised learning based ongradient descent by way of distance metric between object pairs based on all features. A KNN model was explored on all 3 classification tasks using the tabular dataset. The following hyperparameters were varied: rebalancing mode and number of neighbors. Happily, run-times were reasonable to excellent using the KNN approach, and as shown in the associated tuning table, all variations yielded excellent accuracy, precision, and recall for both the superclass and stellar subclass tasks [18]. While both rebalancing methods were excellent for the superclass, there was a noticeable performance degradation using the rarest approach for stars, as compared to smote. Surprisingly, fewer neighbors, as a rule, led to better models. The galaxy subclassification task yielded poor results, with best accuracies around 25%.

**The multiclass logistic regression** model is a supervised learning technique that induces a linear hyperplane separating labeled classes by applying the gradient descent approach to the logistic loss. Nonlinearity is established by way of the softmax function, and assignments are made upon application of a decision rule. In our case, label predictions were assigned based on the class with highest computed probability. The following hyperparameters were varied: rebalancing mode, solver, penalty, C, L1 ratio, and maximum number of iterations [19]. Most run-times lasted less than a minute, and performance metrics were otherwise generally similar to the KNN model. The fastest and most accurate models were those that used SMOTE, applied the lbfgs solver, had no penalty, and had fewer maximum iterations. Accuracies for superclass were achieved upwards of 99% for superclass and over 91% for stellar subclass, but this model fared poorly in predicting galaxy subclass, with the best being less than 23% accurate.

**The Support vector machine classifier** is…Results are…

**The Decision tree classifier**  is…Results are…

**The Random forest classifier** is…Results are…

**The Multilayer Perceptron model** is a predictive model prototypical of feed-forward artificial neural networks…Re**sults are…**

A **feed-forward neural network model** was applied to the image dataset. Results are…

A **resnet model** was applied to the image dataset. Results are…

A **convolutional neural network** (CNN) model was applied to the image dataset. Results are…

The **long short-term memory** model (LSTM) exemplifies recurrent neural network models. A LSTM model was applied to the image dataset. Results are…

**Discussion**

…Summary of train+validation and test metrics for each tuned classification model of each classification task…Summary of conclusions… Limitations… Future Work…

**References**

A - [Pierre Simon Laplace (1749–1827) | Nature](https://www.nature.com/articles/163468a0)

B - [Vera Rubin Will Generate a Mind-Boggling Amount of Data - Universe Today](https://www.universetoday.com/164565/vera-rubin-will-generate-a-mind-boggling-amount-of-data/)

C - Lintott et. All (2008)

D - https://www.sdss.org/science/

E - [Instruments - SDSS](https://www.sdss.org/instruments/)

F - Lupton et. All (2001)

G - Hart et. All (2016)

H - Willett et. All (2013)

**Figures**

1 – Hubble Tuning Form Galaxy Classification Scheme (simplified)

A diagram of a galaxy

Description automatically generated

2 – Hertzsprung-Russel Diagram of Main Sequence Stars based on observational study

A screen shot of a computer screen

Description automatically generated

3 – Morgan-Keenan Color Index Classifications from Stellar Spectra

A chart of different colors

Description automatically generated

4 – Spatial distribution of objects

A diagram of a blue and green object

Description automatically generated

5. Simplified Galaxy Zoo 2 Classifications

A screenshot of a computer

Description automatically generated

6 – Simplified MK Classifications

A white paper with black text

Description automatically generated

7 – Object Superclass Label Counts

A graph with different colored squares

Description automatically generated

8 – Galaxy Subclass Label Counts

A graph of numbers and a bar chart

Description automatically generated with medium confidence

9 – Stellar Subclass Label Counts

A bar graph with numbers

Description automatically generated

10 – Sum\_p

A graph of a number of different colored bars

Description automatically generated

11 – u-r\_p

A graph of different colored bars

Description automatically generated

12 – g-i\_p

A graph with different colored bars

Description automatically generated

13 – r-z\_p

A graph of different colored bars

Description automatically generated

14 – Log Transformed Redshift

A graph of a function

Description automatically generated with medium confidence

15 – Variance-Covariance Heatmap of Selected Tabular Features

A screenshot of a graph

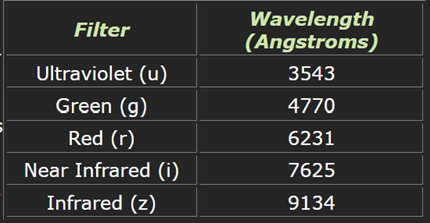
Description automatically generated

16 – Pairplot of selected features

A collage of images of different colored dots

Description automatically generated

17 – [SDSS Color Filters](https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp)



Y – superclass PC

A graph with a line

Description automatically generated

Y – superclass PC

A graph with blue squares

Description automatically generated

Y – superclass PC

A graph with blue squares

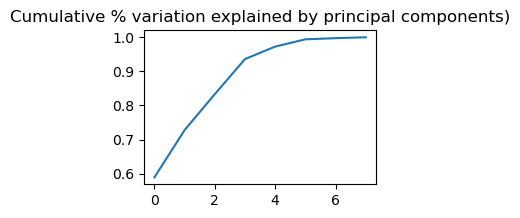
Description automatically generated with medium confidence

Y – Superclass PC

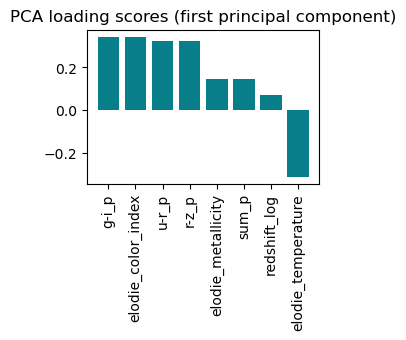
A graph of different colored objects

Description automatically generated with medium confidence

X – Star PC



X – Star PC



X - Star PC

A graph with blue and green squares

Description automatically generated with medium confidence

X – Star PC

A diagram of a scatter diagram

Description automatically generated with medium confidence

18 – KNN Tuning Table



19 – Logistic



20 – SVM



21 – schematic of GZ2 decision tree

A diagram of a spiral pattern

Description automatically generated with medium confidence

22 – tabular preprocessing diagram

A diagram of a process

Description automatically generated

23 – Decision Tree



24 – Random Forest

