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 J. Spencer Morris

Date: 14 December 2023

GitHub Repository: https://github.com/jspencermorris/datasci207_project

Contents:

- Motivation
- Objectives
- Classification Schema
- SDSS Project
- Galaxy Zoo Project
- Datasets
- Evaluation Methodology
- Tabular Preprocessing
- Image Preprocessing
- Model Details
- Classification Results
- Discussion

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.[1] 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.[2]

This paper provides a demonstration of celestial object classification using a suite of machine learning models as applied to data of various shapes originating 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 the Morgan-Keenan 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 Hubble 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 ~10^{24} 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 ~10^{12} solar masses and lengths ~10^{20} 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 ~10^{12} 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 [Fig. 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 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 ~10 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 ~10^{11} 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 ~10^{10} 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, and is visualized as a 'main sequence' along the primary dimensions of luminosity and color index in a Hertzprung-Russell Diagram [Fig. 2]. A star's color index, therefore, is its foundational classification along the main sequence, and has been formalized using the Morgan-Keenan (MK) system. 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) [Fig. 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 ~10 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.[3] 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.[4] 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 accession 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) [Fig. 4], and two double spectrographs (with a combined wavelength range of 3600 to 10400 Å) paired to CCD's.[5] Even at inception, therefore, SDSS generated 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.[6]

Galaxy Zoo Project

The early SDSS Legacy study ultimately characterized ~10^6 galaxies via photometry, spectroscopy, and imaging, enabling galaxy studies of unprecedented scope. Its image catalog 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 ~10^7 individual classifications by ~10^5 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 [Fig. 5].[7] 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.[8]

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, 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.[9] 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 directory contains the SDSS SQL queries. Finally, all finalized work is contained in the /reports directory.

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. All run-times for each tabular model fit were recorded after a memory refresh. Modeling of the image data was conducted locally on a machine with an AMD Ryzen 7 7800X3D 8-Core Processor and 32 GB of DDR5 RAM, with the GPU not utilized.

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.

For all models, we investigated accuracy (a measure of the percent of correctly classified objects), precision (a measure of the percent of true positive predictions among all positive predictions), recall (a measure of the percent of true positive predictions among actual positives), and F-1 score (a measure that reflects the harmonic mean of precision and recall) metrics. Following all tabular preprocessing, our baseline accuracies for balanced train sets were ~33% for superclass, 20% for stellar subclass, and 10% for galactic subclass. F-1 was of primary interest in evaluating model performance on validation sets during tuning. Since all tasks were single-label multi-class classifications, we made extensive use of confusion matrices to visualize subclass ground-truths and predicted label assignments.

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 from multiple classification variables were not contradictory and the object was flagged by SDSS as 'clean' (i.e. no documented artifacts or errors during data acquisition). The three labels available for the superclass classification task were: galaxy, quasar, star [Table 1]. 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 [Fig. 6].

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 an MK label. We considered dropping outliers based on Tukey or Mahanabolis measures but opted to include such data to retain more object 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. Stellar subclassifications were mapped to a reduced set of 9 labels in line with the simplified MK system [Table 2], and galaxy subclassifications were mapped to a reduced set of 14 labels in line with the simplified Hubble Sequence [Table 3].

Even after label engineering, the final label distributions were extremely unbalanced [Figs. 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, an unbalanced training dataset does present a challenge. Machine learning algorithms are most performant when training sets include equally spaced labels (in the case of regression) or equally represented labels (in the case of classification). The unbalanced nature of the dataset was addressed and is discussed later in this paper.

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 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 [Figs. 10,11,12,13].

Finally, the redshift value, as determined from each object's spectrum, was selected. Redshift, when log transformed, was determined to exceptional discriminatory power across superclass labels [Fig. 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 [Fig. 15], as is a pairplot with indication of superclass [Fig. 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 further reduced the number of evaluated star labels to 5 and the number of evaluated galaxy labels to 10. Our baseline accuracies for the train sets were therefore ~33% for superclass, ~20% for stellar subclass, and ~10% for galactic subclass. 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 which, 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 subclass, 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 [Fig. 17], and the redshift and photometric

sum dominate those variable loadings [Figs. 18,19]. Projection of the objects along the first 3 components shows good separation by superclass [Fig. 20]. For the stellar feature set, ~85% of the underlying variation was explained using just two principal components [Fig. 21], where the temperature, photometric sum, metallicity, and redshift had the highest loadings [Figs. 22,23]. A scatter plot of the first two principal components allows for visualization of how this unsupervised learning model can be useful for stellar classification, and is reminiscent of the Hertzprung-Russell Diagram [Fig. 24].

The diagram of the preprocessing pipeline summarizes the overall approach used for the tabular dataset [Fig. 25].

Image Preprocessing

After sampling for the images used in the model, some preparation was required to input the data into the image classification model. The models used for classifying galaxy images were Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Feed-Forward Neural Networks (FNN), and finally a Resnet Model (simply put, a pretrained CNN model). These models require inputs to be standardized, and data augmentation to improve the generalization of the outputs of the model.

Initially, each image is loaded from a specified directory, and resized. Resnet models have a defined architecture, built on trained images of sizes 224x224; for that reason, images were resized to 224x224 to maintain this consistency in the model's inputs. Corresponding labels were extracted for each loaded image as the identifier for the galaxy classification. Images were initially loaded in at 424x424, however they were cropped down to 224x224. It was done this way as opposed to immediately loading the image in as a 224x224 to ensure that the central part of the image was captured since the galaxies are centered in the images.

Following this, some adjustments were made in the contrast and brightness of the images. According to the literature, some defining features of the galaxy are in the colors, for example, spiral galaxies are known to have a blueish-white color to them.[10] Ensuring that these colors can be contrasted from others is essential to get an appropriate classification. At this point, each processed image is then transformed into array format in compliance with the requirements for deep learning frameworks. This concluded the initial image preprocessing.

In the second round of preprocessing, the images were randomly shuffled and split into training, validation and testing sets. The distribution of train, validation, and testing sets were 60%, 20%, 20%. The data was then normalized, scaling RGB values from 0-255 to [0,1] to standardize the input values. Training set image brightness and contrast levels were further adjusted. Following this, the images were randomly flipped, making the model invariant to the direction of the image feature orientation. The original training set was then concatenated along with the augmented training set to provide the model with a larger set to train with, thus reducing the risk of overfitting. A final shuffle was then performed to randomly distribute the data between the original data and augmented data.

For the full model training, the 10 most common classes were chosen and randomly sampled to handle the imbalances in the dataset and cut down the amount of data to make training more manageable. 1600 images were sampled in each class, and through the 60% training split, 9600 images were randomly chosen from the various classes before being doubled in augmentation for a final total of

19200 training images. From the remaining images, there were 3200 validation data points, and 3200 testing data points. The final distribution of the classes in the training set can be seen in [Fig. 26].

Examples of typical images ranged in color and morphology [Fig. 27].

Model Details

The simple linear regression model describes a linear relationship between a metric independent variable and a metric dependent variable, and can be solved either analytically with ordinary least-squares or iteratively with gradient descent (whereby the minimized objective function is the mean-squared error). We modeled the Elodie Color Index as a linear function of temperature (which was log-transformed) and observed slight departures from linearity for the hottest stars [Fig. 28]. The resulting model's high coefficient of determination of ~ 0.98 for the test data demonstrates the SDSS-sourced temperatures are themselves likely modeled from the Elodie Color Index (which is itself determined from the difference in brightness of two spectral bands).

The **k-means clustering** model is an unsupervised learning model of hypershperical clusters of objects based on minimization of pairwise Euclidean distances, is prototypical of clustering, is computationally efficient, and requires an a priori assumption of the number of clusters within the dataset. K-means models were explored for the three transformation sets described above (X0, X3, and X7) and for both feature-sets (which contained 5- and 8- features). The following hyperparameters were varied: initialization method, number of clusters, number of centroid seeds, and the maximum number of iterations. Frustratingly, none of the clustering models well-described any of the modeled datasets, and we observed the tendency of galaxy objects to be assigned to clusters associated with quasars. An elbow plot for the first feature-set (applicable to superclass) was predicted to have a strong inflection point at 3 clusters, and the lack thereof illustrates the unsuitability of this model [Fig. 29]. We suspect that the model's failings are due to its requirement for feature-space hypersphericity.

The **Gaussian mixture model** (GMM) is similar to k-means in that it is an unsupervised clustering model but differs in that clusters are described as overlapping probabilistic distributions and also allows for multi-cluster assignment to objects. We did not explore this model for our dataset, but we suspect its relaxed requirement for cluster shape will lead to superior performance as compared to k-means.

The **naïve Bayes model** is a supervised learning model that classifies objects based on application of Bayes' Theorem, where the observed feature values associated with each label inform label assignments. We did not explore this model for our dataset, but reserve it as a future exercise.

The **k-nearest neighbors** classification model exemplifies supervised learning in that an object's predicted label is based on a majority-vote of the known labels of nearest neighbors (as computed using pair-wise distance metrics in the feature-space). 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 [Table 4]. We observed that models trained on the oversampled set were slightly more predictive (though had run-times about 3 times longer) than those trained on the undersampled set, and that fewer neighbors led to better models for superclass and stars (while more neighbors was better for galaxies). Using our best

hyperparameters, validation F-1 for superclass approached ~100%, for stellar subclass was ~98%, and for galaxy subclass was ~27%.

The **logistic regression** classifier model is a supervised learning model that induces a linear decision boundary separating labeled classes by applying the gradient descent approach to the logistic loss. For all of our multiclass tasks, nonlinearity was established using a softmax function, and label predictions were assigned based on the class with highest computed probability using the one-vs-rest ('ovr') hyperparameter. The following hyperparameters were varied during tuning: rebalancing mode, solver, penalty (which was either none, L2 only, or both L1 and L2), C (the strength of the L2 regularizer), L1 ratio (to control relative strength of L1 and L2 regularizer terms), and the maximum number of iterations [Table 5]. Most run-times lasted less than a minute, and performance metrics were similar for all hyperparameter combinations and comparable to the KNN model. We found that oversampling did not perform significantly better than undersampling for any cases (although the impact to run-time was fairly small), the saga solver (which included both L1 and L2 regularizers) did not significantly improve predictive performance but did significantly degrade run-time performance, it was not necessary to include an L2 regularization term, and 50 iterations was sufficient for good predictions. Using our best hyperparameters, validation F-1 for superclass approached ~100%, for stellar subclass was ~94%, and for galaxy subclass was ~22%.

The **support vector machine** (SVM) classifier model is a supervised learning model that maximizes the margins between a class-separating hyperplane and the most extreme class examples. The 'kernel trick', which transforms a low-dimensional nonlinear hyperplane into a high-dimensional linear hyperplane, enables improved classification performance for the SVM model, and all of our models employed the radial basis function ('rbf') for kernelization. The following hyperparameters were varied during training: rebalancing mode, the L2 regularization parameter C, and the number of maximum iterations for the solver [Table 6]. We found undersampling typically had better validation performance than oversampling, a regularization parameter of 1 was best, and that there was a marked improvement when increasing the number of maximum iterations from 10 to 100, though this ~doubled run-times. Using these hyperparameters, validation F-1 for superclass was ~96%, for stellar subclass for ~91%, and for galaxy subclass was ~11%.

The **decision tree** classifier model is a supervised learning model that repeatedly splits objects based on the feature that results in the largest information gain (IG) and is therefore capable of building complex decision boundaries since this approach essentially divides the feature space into rectangles at each layer. The following hyperparameters were varied during training: rebalancing mode, IG criterion, minimum samples required to split an internal node, and the maximum depth of the tree [Table 7]. We observed that undersampled train sets led to comparable predictive performance with oversampled train sets for superclass and star classification tasks, although oversampling was slightly helpful for the galaxy classification task. Entropy was observed to be a superior IG compared to Gini (especially for stars). Varying the minimum samples required for a split did not matter much, but we found that a maximum depth of 2 was insufficient for the stellar and galaxy subclassification tasks (which is not surprising given the number of subclasses is greater than 3). Using our best hyperparameters, validation F-1 for superclass approached ~100%, for stellar subclass approached ~100%, and for galaxy subclass was ~27%. Decision Trees are particularly interpretable classifiers since they enable visualization of the most informative features. For example, the decision tree schematic for our best superclass model shows that an initial split of objects with log redshifts less than

-1 is essentially 100% predictive of stars vs. galaxies/quasars, and that a second split with a log redshift cutoff of 0.4 is essentially 100% predictive of galaxies vs. stars [Fig. 30].

The **random forest** classifier model is a supervised learning model that makes use of ensembling to randomly generate many decision trees prior to holding a majority vote for final predictions. The following hyperparameters were varied during training: rebalancing mode, bootstrap sampling, and number of estimators [Table 8]. For all classification tasks, oversampling led to slightly better predictive performance than undersampling, but at a significant cost of ~10x longer run-times, and we therefore considered oversampling preferable only for the galaxy classification task. Bootstrap sampling was associated with slightly better results, and we found 10 estimators sufficient for the superclass and star tasks (whereas 50 estimators led to better outcomes for the galaxy task). Using our best hyperparameters, validation F-1 for superclass approached ~100%, for stellar subclass approached ~100%, and for galaxy subclass was ~28%.

A **feedforward neural network** (FFNN) model is a supervised learning model that leverages backpropogation and a series of densely connected hidden layers. The final model studied for the tabular dataset was a simple FFNN utilizing the Adam optimizer that incorporated two hidden layers with 512 and 128 neurons (and therefore had 69,123 total trainable parameters for the 5-variable feature-set and had 70,917 parameters for the 8-variable feature-set used for stellar classifications). All models had a batch size of 2048 and were trained for 5 epochs to monitor achievement of overfitting. The following hyperparameters were varied: rebalancing mode, activation function, and learning rate [Table 9]. Predictive performance metrics were investigated after the final epoch, and we observed convergence after 5 epochs only for most models (with the exception of undersampled stellar and galaxy models). Oversampling was demonstrated to be necessary for best predictive performance for the stellar and galaxy subclass tasks. For all models, relu activation yielded somewhat faster run-times than tanh, and a learning rate of 0.1 was best. Using our best hyperparameters, validation F-1 for superclass approached ~100%, for stellar subclass was ~98%, and for galaxy subclass was ~28%.

A second **FFNN** model was also utilized as a baseline for galaxy subclassification from the images data. The model is comprised of 4 dense layers that feed into each other, each one followed by batch normalization. At the end is a single dense layer with SoftMax activation. Each hidden layer is using Relu activation. Of all the image classification models, this model had the lowest predictive performance, becoming heavily overfit after 50 epochs with a validation accuracy ~20% [Fig. 31]. As is evident from a confusion matrix, the model tended to heavily over-predict on a small subset of the classes rather than learn meaningful features on all classes [Fig. 32].

The RestNet50 **res-net** model was implemented to demonstrate **transfer learning**, further training a pre-trained model used commonly for image recognition tasks on the galaxy dataset. In this case, the res-net was fed directly into a dense layer before the final layer with soft-max activation. The res-net model was originally trained on 224 x 224 x 3-dimensional images, the same as the images it is training and testing on. The res-net performed very well, but was unable to improve past the first few epochs, quickly becoming overfit at 51% accuracy [Fig. 33].

The **convolutional neural network** (CNN) model begins with a layer that randomly rotates the image to attempt to address overfitting and is followed by 3 series of convolutional layers. Each series is comprised of a conv2D, a MaxPool2D, and a batch normalization. They then feed directly into the next series, until they are flattened after the third convolution. After flattening, they go through 2 dense layers, each followed by batch normalization. They then feed into the final SoftMax layer. All

convolutional layers are using Relu activations. The CNN, which consisted of 5,376,0970 parameters, performed very well, reaching 51% accuracy [Fig. 34]. As the CNN is explicitly trained on galaxies, it is better able to understand features unique to this dataset.

The **long short-term memory** (LSTM) model exemplifies recurrent neural network models. The LSTM takes the architecture of the previous CNN through the flattening layer, but time distributes the model in 32 batches. The output of the time distributed and flattened features go into the LSTM layer, which then goes into the dense SoftMax. The LSTM model performed well, after 50 epochs reaching 44% accuracy on validation and test data [Fig. 35]. The reduced performance compared to a standard CNN is not unusual, as the applications of an recurrent network does not specifically benefit image classification in the same way it would a natural time series application, such as video classification.

Discussion

In summary, 12 models were explored using 2 datasets and applied to 3 classification tasks, resulting in full specification of 28 tuned models [Table 10]. Principal component analysis was used to generate feature sets of reduced dimensionality and demonstrate structural clusters within the tabular dataset. Simple linear regression was used to demonstrate the relationship between Elodie Color Index and temperature of stars. K-means was attempted to evaluate clustered structures. 3 single-label multiclass classification tasks (superclass, stars, and galaxies) were performed on the tabular dataset using k-nearest neighbors, logistic regression, support vector machine, decision tree, random forest, and feedforward neural network models. Finally, single-label multiclass classification of galaxies was performed on the imagery dataset using a feedforward neural network, a pre-trained res-net, a convolutional neural network, and a long short-term memory model. Final summaries of the test performance metrics of the best tuned models for each classification task [Tables 11,12,13].

All models had excellent predictive performance for the superclass classification task. We found that random forest classification yielded the highest F1, at ~99.94%, and that run-time performance was very good at just 26 seconds. A confusion matrix demonstrates that only a few confusions were made between galaxies and stars, with nearly perfect quasar prediction. [Fig. 36].

For the stellar subclass classification task, most models had fairly good predictive performance. The decision tree model yielded the best F1, at ~99.72%, and furthermore had excellent run-time performance of just 2 seconds. Of this model's few misclassifications, most were associated with G-type stars improperly labeled as either K or F, which is not unsurprising since these star classes are very similar in color. Fig. 37].

The predictive performance of all models was substantially lower for the galaxy subclass classification task, but we are not surprised that the tabular dataset yielded somewhat lower results (since those features necessarily contain limited information pertinent to morphological classification). Even still, the best F1 score based on tabular data was ~34.12%, which is still substantially better than our baseline. A confusion matrix demonstrates that misclassifications were extensive and often of subclasses of similar morphologies. Fig. 38].

For galaxy subclass classification using imagery, the Convolutional Neural Network performed best. Its predictive performance often spiked above that of the res-net while surpassing the LSTM and FFNN models. In comparison to the res-net, the CNN took significantly less time to train, and showed potential for further improvement [Fig. 39]. The res-net did show much less variation in validation however, due

to having better established weights in the model. The CNN predicted best on classes 3 and 6, but struggled with 0 and 9, and heavily misclassified many class 9 images as class 5 [Fig. 40].

One of the limitations of the image recognition models was the excessive training time which limited the capabilities of excessive testing. The res-net presented the greatest problem, needing 224x224 for proper performance but taking more time to train than any other model. The CNN parameters were tuned meanwhile on 128x128 images to speed up processing time, as kernel size and stride size were consistently adjusted for better performance, as well as the number of convolutional layers, filter size, and whether to use drop out or batch normalization. Image quality was also a concern, as some of the images from the Galaxy Zoo 2 dataset appear to be distorted or lower quality.

We have many ideas for further improving modeling of cosmic objects. Our data acquisition (including queried features, filtering, and preprocessing) necessarily limited our conclusions to objects that are relatively large and bright, and we think it would be interesting to test the impact of relaxing these constraints on predictive performance. More analysis of the k-means clustering model is clearly warranted given its apparent poor performance, and we believe exploring additional transformed feature-sets and decision-boundary plotting may provide greater insight into why, and we also believe a gaussian mixture model may yield better results. Our approach to hyperparameter tuning was time-consuming and could be improved through automation via a tuning scheduler and/or a grid search; this would furthermore enable much greater variation of hyperparameter values for more extensive model investigations. We are also interested in applying k-fold validation for more robust predictive performance metrics, which were especially noisy for models applied to the imagery dataset, as well as visualizing the accuracy/precision tradeoffs of our models using ROC curves. Given our primary interest in classification of galaxy morphologies, we are confident that longer training times of our CNN model will lead to increased predictive performance, and furthermore would have liked to examine additional labels (such as binary GZ2 labels associated with number of spiral arms, etc.). We believe utilizing a cloud-based service such as Google Colab or Amazon Web Services may yield improved run-times. As this is an area of active research, we would have liked to apply additional models (including a naïve Bayes model and a transformer model). We hope in the future to compare our approaches to other recent studies for galaxy morphology.[10,11] Finally, we are interested in applying our models to non-SDSS datasets to study their generality and applicability as components of preprocessing pipelines in next-generation astronomical surveys.

Figures

Figure 1 – Simplified Hubble Tuning Fork for galaxy classification (source: wikipedia)

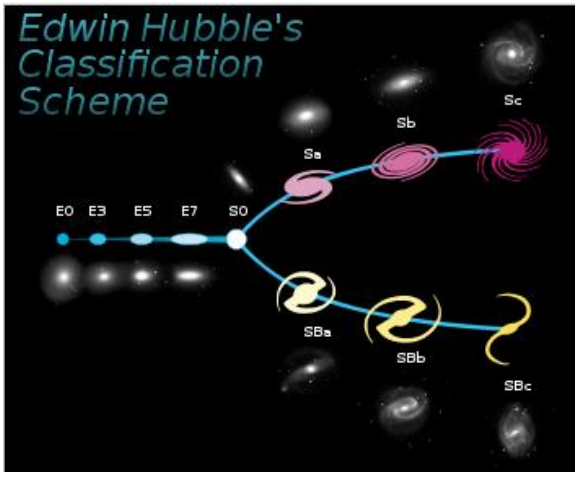


Figure 2 – Hertzsprung-Russel Diagram of main sequence stars based on observational study (source: wikipedia)

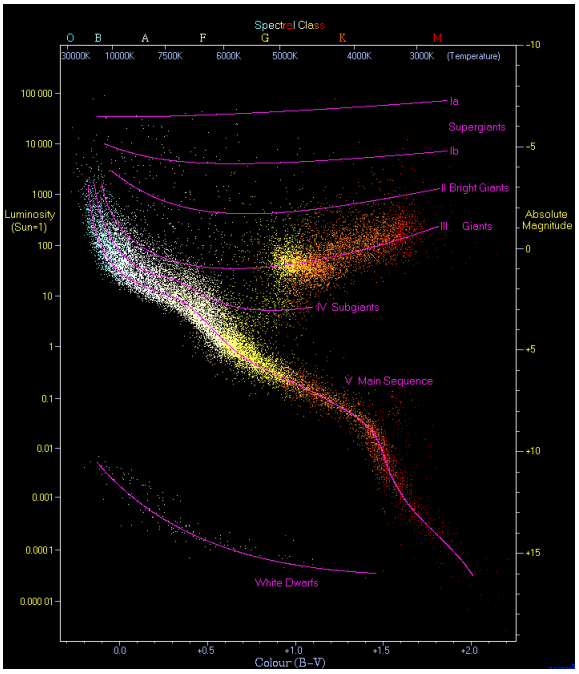


Figure 3 – Morgan-Keenan Color Index classifications from stellar spectra (source: wikipedia)

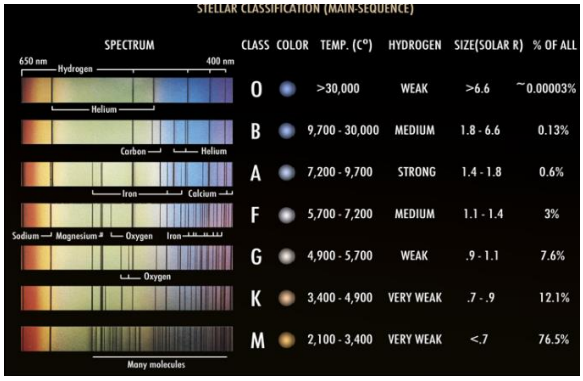


Figure 4 – Most responsive wavelengths for SDSS Color Filters (source: https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp)

Filter	Wavelength (Angstroms)
Ultraviolet (u)	3543
Green (g)	4770
Red (r)	6231
Near Infrared (i)	7625
Infrared (z)	9134

Figure 5 – Schematic of GZ2 decision tree (source: Lintott et. al. [2008])

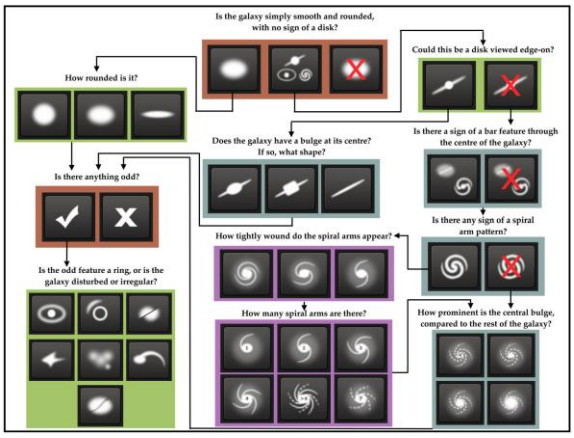


Figure 6 – Spatial distribution of examined objects

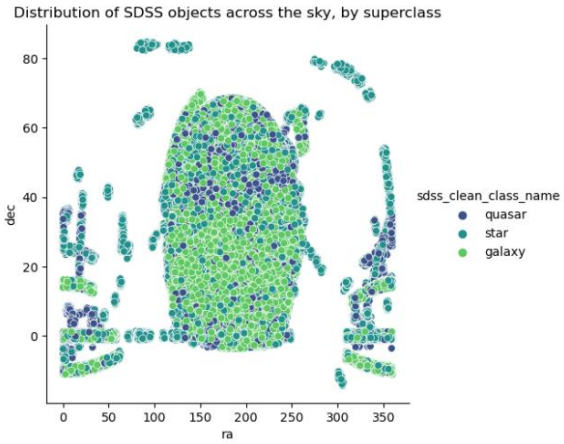


Figure 7 – Superclass label counts after high-fidelity filtering

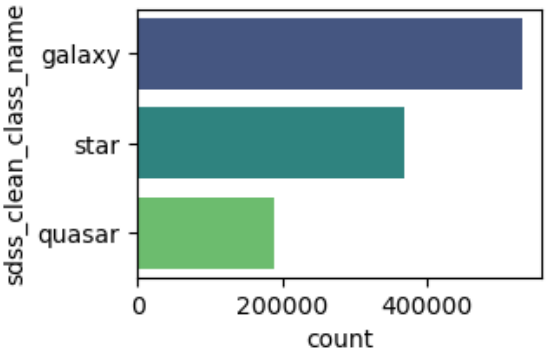


Figure 8 – Galaxy subclass label counts after label engineering

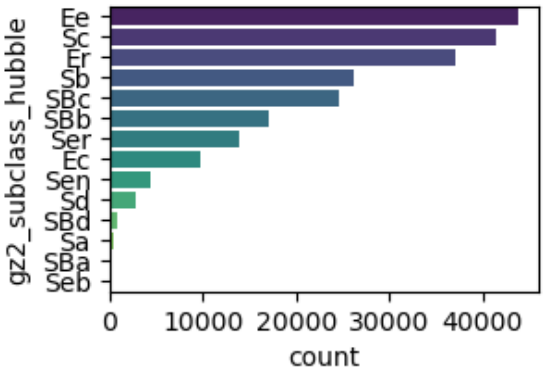


Figure 9 – Stellar subclass label counts after label engineering

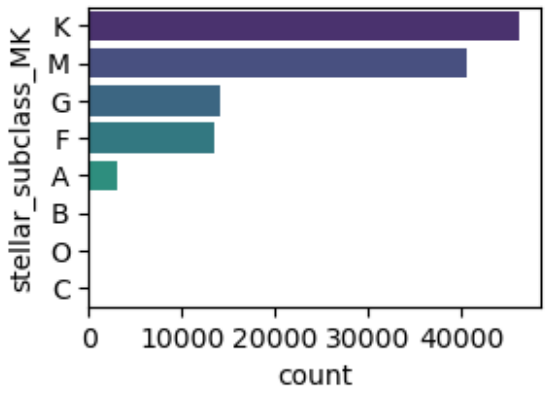


Figure 10 – Histogram of the sum of all photometric values

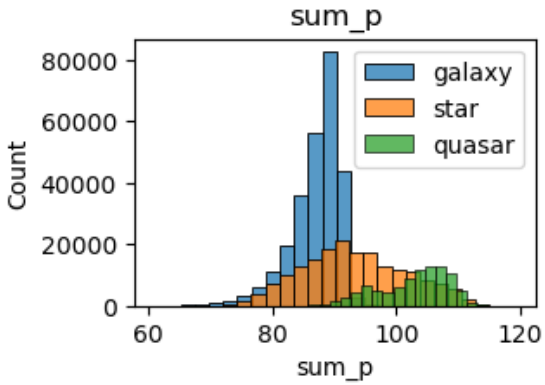


Figure 11 – Histogram of the difference between ultraviolet and red photometric values

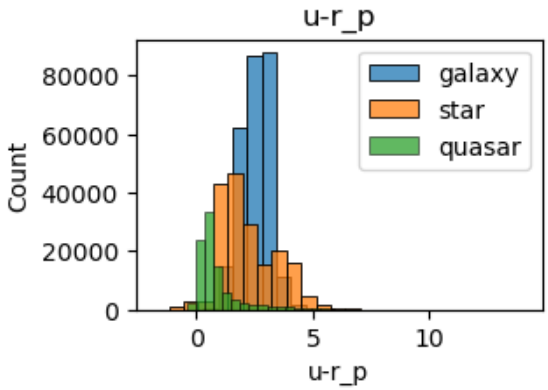


Figure 12 – Histogram of the difference between green and infrared photometric values

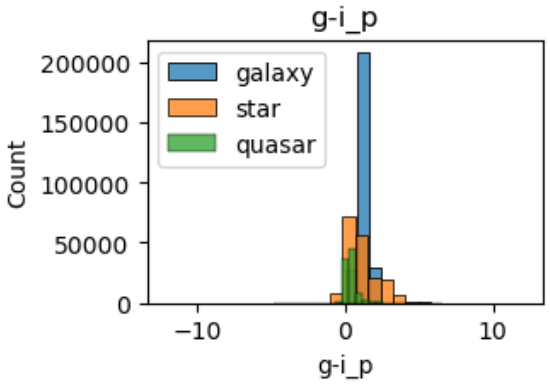


Figure 13 – Histogram of the difference between red and deep-infrared photometric values

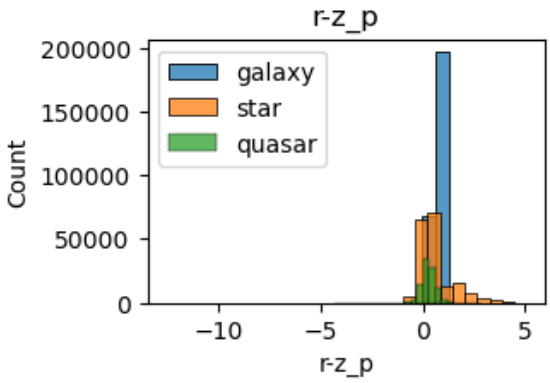


Figure 14 – Histogram of logarithm of spectroscopic redshift

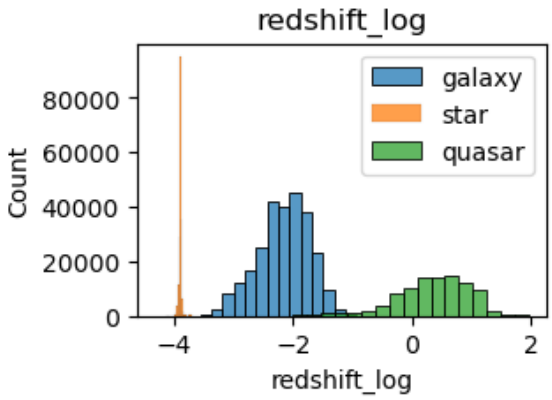


Figure 15 - Variance-covariance heatmap of selected tabular features

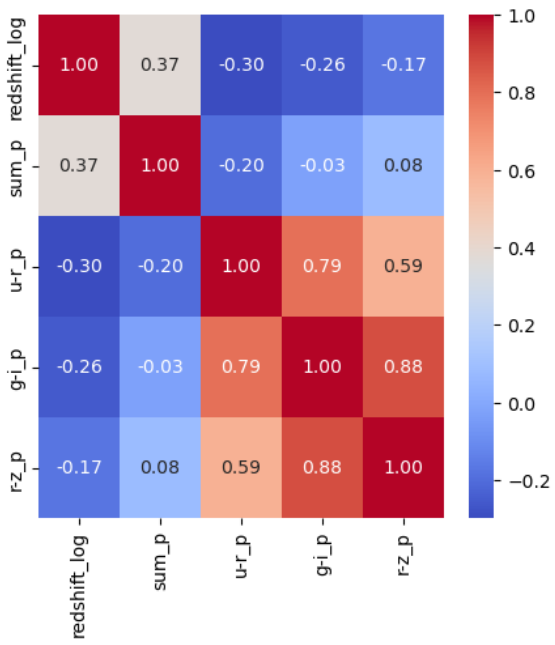


Figure 16 – Pair-plot of selected tabular features

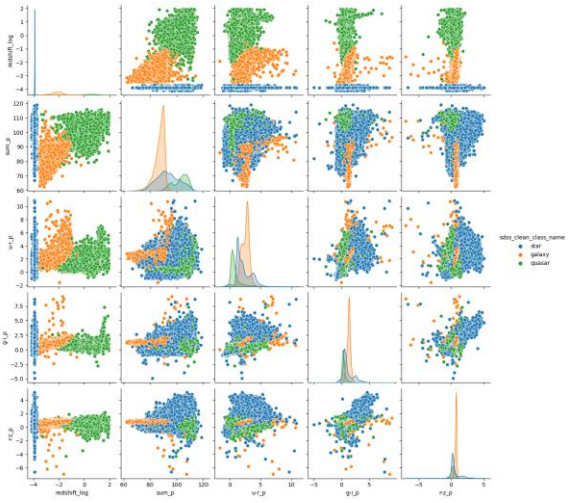


Figure 17 – First feature set's elbow plot of principal components

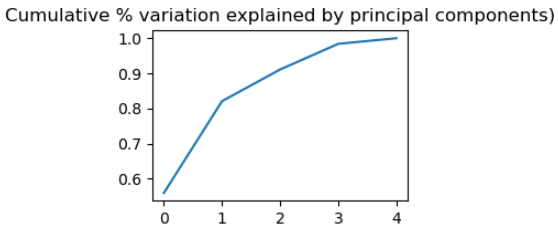


Figure 18 – First feature set's loading scores for first principal component

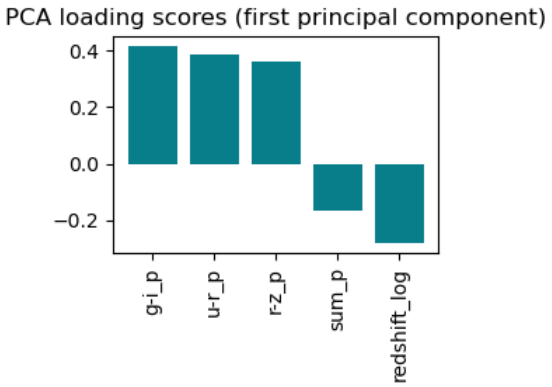


Figure 19 – First feature set's loading scores for second principal component

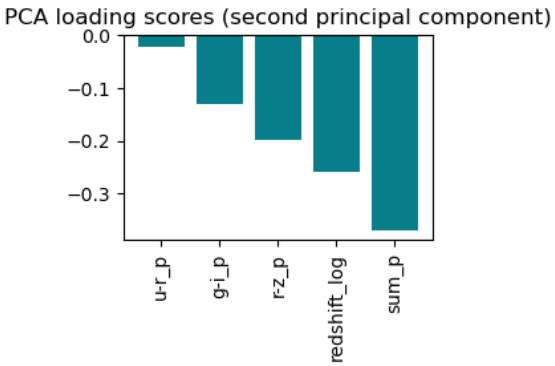


Figure 20 – First feature set's scatter along first three principal components, with indication of superclass

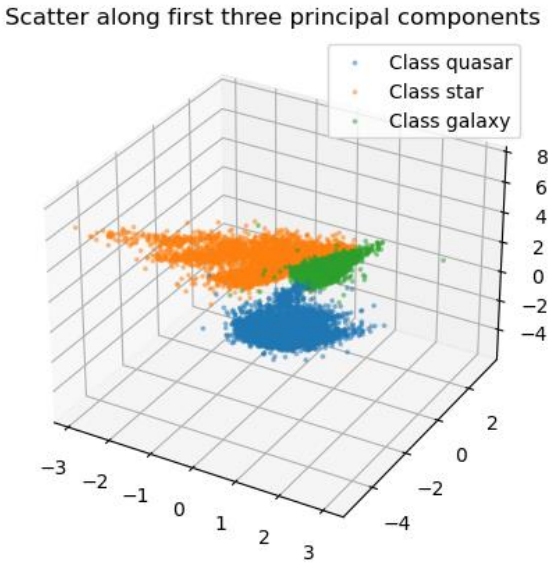


Figure 21 – Second feature set's elbow plot of principal components

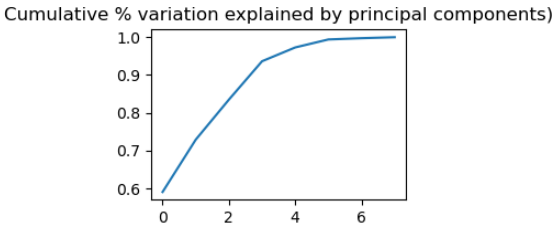


Figure 22 – Second feature set's loading scores for first principal component

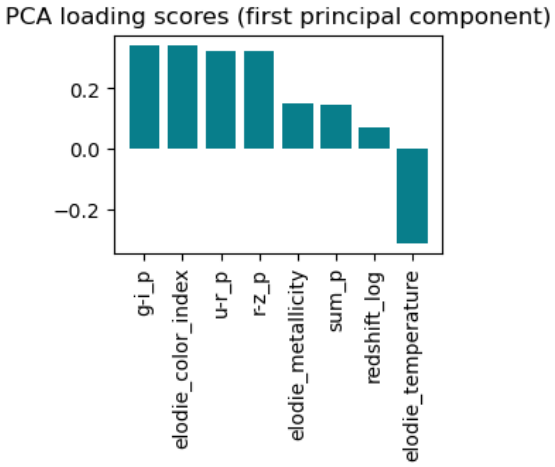


Figure 23 – Second feature set's loading scores for second principal component

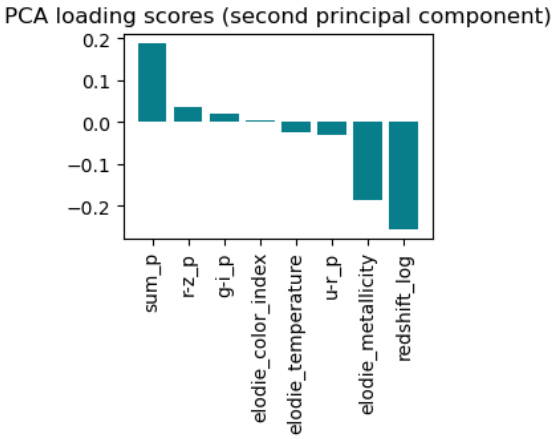


Figure 24 – Second feature set's scatter along first three principal components, with indication of superclass

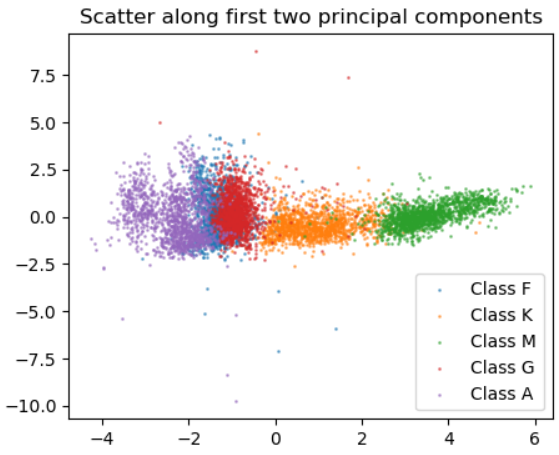


Figure 25 – Schematic of preprocessing pipeline for tabular data

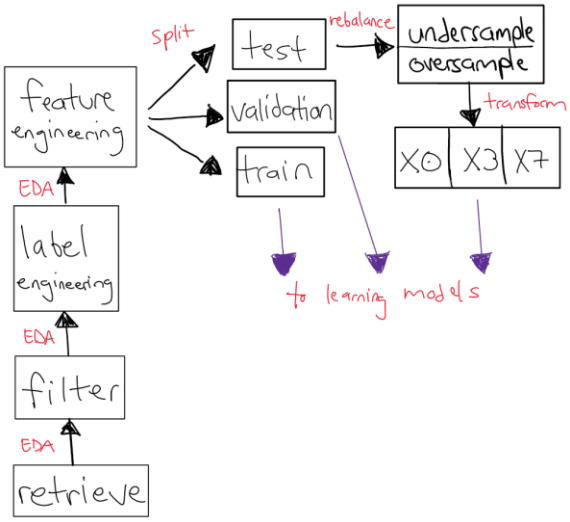


Figure 26 – Distribution of training data classes for image recognition

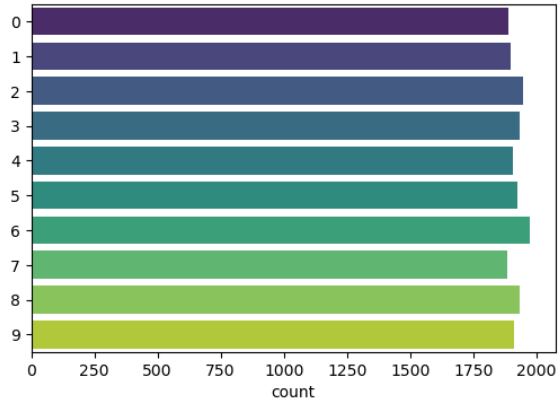


Figure 27 – Examples of galaxy images after preprocessing

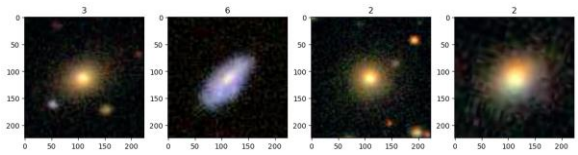


Figure 28 – Linear model of temperature and Elodie Color Index

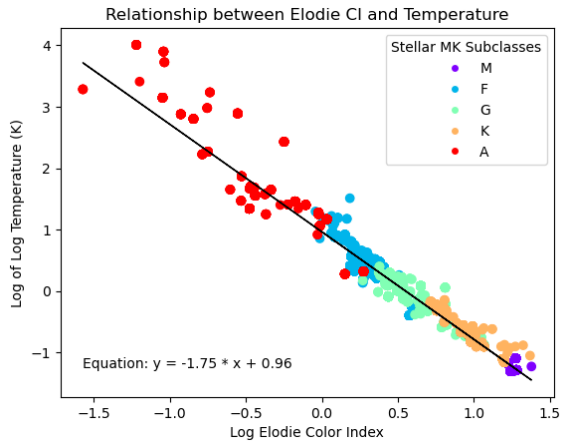


Figure 29 – Elbow plot for superclass dataset, showing poor agreement with a priori cluster number

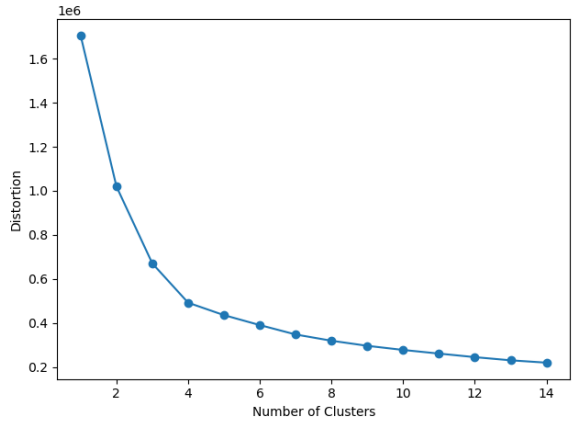


Figure 30 – Decision tree schematic for superclass task

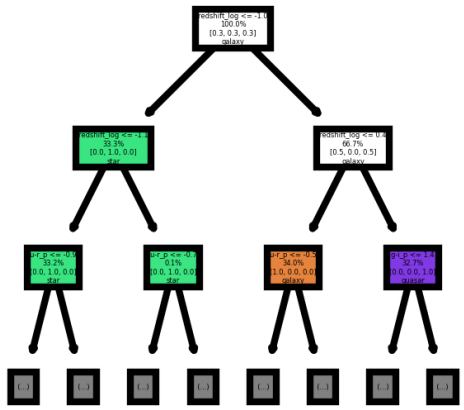


Figure 31 – Performance history for galaxy subclassification task of images using FFNN

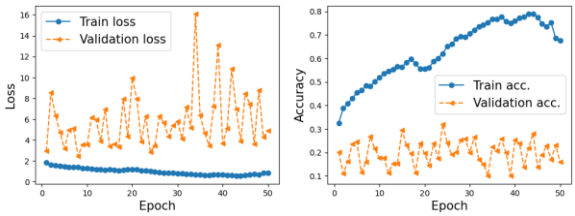


Figure 32 – Confusion matrix for galaxy subclassification task of images using FFNN

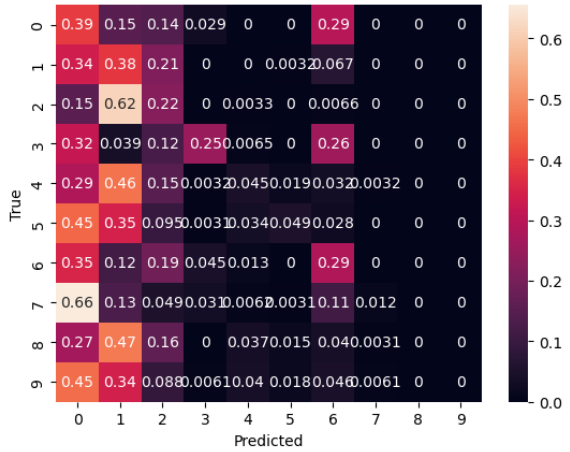


Figure 33 – Performance history for galaxy subclassification task of images using res-net

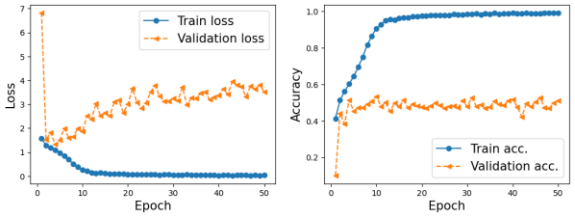


Figure 34 – – Performance history for galaxy subclassification task of images using CNN

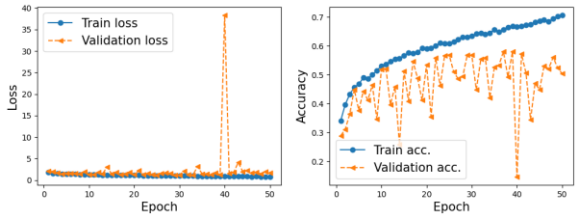


Figure 35 – Performance history for galaxy subclassification task of images using LSTM

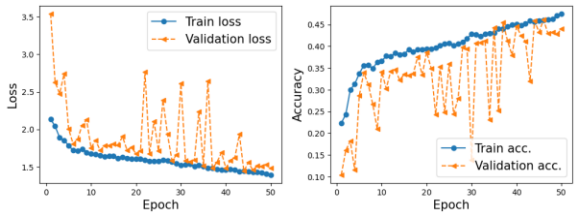


Figure 36 – Confusion matrix for superclass classification task of tabular data using RF

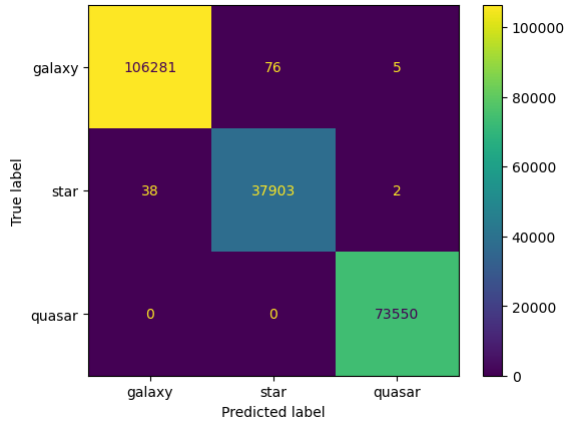


Figure 37 – Confusion matrix for star subclassification task of tabular data using DT

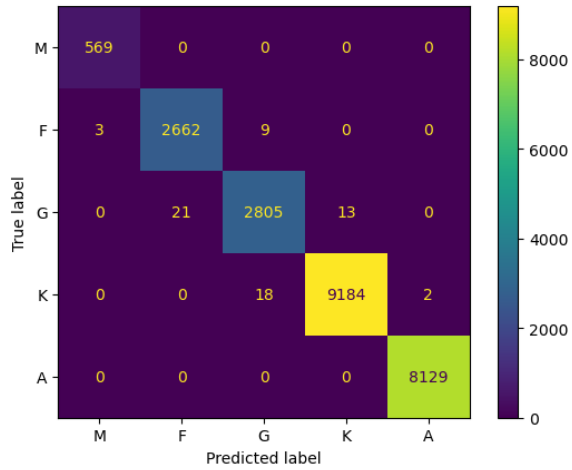


Figure 38 – Confusion matrix for galaxy subclassification task of tabular data using FFNN

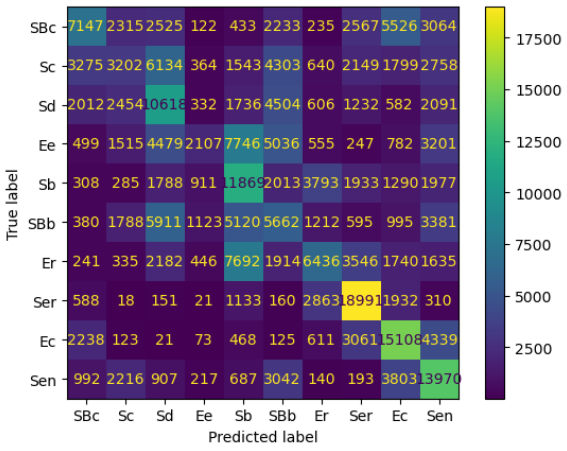


Figure 39 – Performance history for galaxy subclassification task of images comparing all models

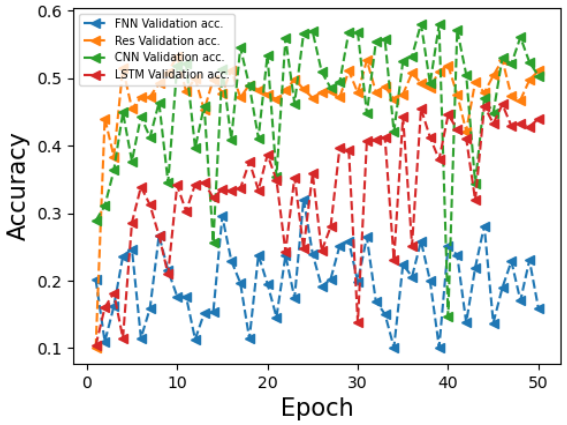
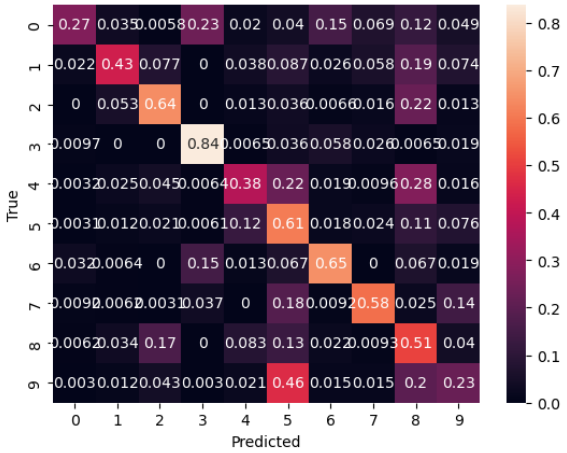


Figure 40 – Confusion matrix for galaxy subclassification task of imagery data using CNN



Tables

Table 1 – Superclass classification labels

Label
galaxy
quasar
star

Table 2 – Stellar subclassification labels

Label	Description
O	blue
B	bluish white
A	white
F	yellowish white
G	yellowish white
K	light orange
M	orangeish red
D	white dwarf
C	carbon star
d	cool (red or brown) dwarf

Table 3 – Galaxy subclassification labels

Label	Description
Er	elliptical with low eccentricity (round)
Ee	elliptical with intermediate eccentricity
Ec	elliptical with high eccentricity (cigar-shaped)
Sa	spiral with large bulge
Sb	spiral with medium bulge
Sc	spiral with small bulge
Sd	spiral with no bulge
SBa	barred-spiral with large bulge
SBb	barred-spiral with medium bulge
SBc	barred-spiral with small bulge
SBd	barred-spiral with no bulge
Ser	edge-on spiral with round bulge
Seb	edge-on spiral with boxy bulge
Sen	edge-on spiral with no bulge

Table 4 – Tuning table for k-nearest neighbors (tabular dataset)

							Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	N Neighbors	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
					5	140	0.9997	0.9997	0.9997	0.9997	0.9994	0.9994	0.9994	0.9994
					20	200	0.9994	0.9994	0.9994	0.9994	0.9992	0.9992	0.9992	0.9992
				Smote	100	308	0.9991	0.9991	0.9991	0.9991	0.9987	0.9987	0.9987	0.9987
					5	50	0.9995	0.9995	0.9995	0.9995	0.9992	0.9992	0.9992	0.9992
					20	63	0.9992	0.9992	0.9992	0.9992	0.999	0.999	0.999	0.999
			Superclass	Rarest	100	98	0.9987	0.9987	0.9987	0.9987	0.9984	0.9984	0.9984	0.9984
					5	16	0.991	0.991	0.991	0.991	0.9809	0.9809	0.9809	0.9809
					20	21	0.9797	0.9798	0.9897	0.9897	0.9751	0.9754	0.9751	0.9751
				Smote	100	33	0.9628	0.9634	0.9628	0.9625	0.9643	0.9643	0.9657	0.9643
					5	4	0.9668	0.9667	0.9668	0.9666	0.9614	0.9622	0.9614	0.9614
					20	4	0.9445	0.9449	0.9447	0.9442	0.9517	0.9536	0.9517	0.9517
			Stars	Rarest	100	4	0.9147	0.9171	0.9147	0.9133	0.9303	0.9343	0.9303	0.9398
					5	31	0.7005	0.6917	0.7005	0.6905	0.2376	0.2728	0.2376	0.2435
					20	40	0.535	0.5157	0.535	0.5176	0.2501	0.3055	0.2501	0.2567
				Smote	100	65	0.4258	0.4065	0.4258	0.4064	0.2641	0.33	0.2651	0.2665
					5	6	0.4993	0.5043	0.4993	0.4962	0.2409	0.2718	0.2409	0.24429
					20	7	0.4047	0.398	0.4047	0.3985	0.2666	0.3076	0.2666	0.2711
KNN	X0	1000	Galaxies	Rarest	100	9	0.3667	0.3556	0.3667	0.351	0.2716	0.3223	0.2716	0.2728

Table 5 – Tuning table for logistic regression (tabular dataset)

												Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	Solver	Penalty	C	L1 Ratio	Max Iterations	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
									50	34	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992
						None	1		5000	31	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992
									50	31	0.997	0.997	0.997	0.997	0.9966	0.9966	0.9966	0.9966
							0.1		5000	31	0.997	0.997	0.997	0.997	0.9966	0.9966	0.9966	0.9966
									50	30	0.9989	0.9989	0.9989	0.9989	0.9988	0.9988	0.9988	0.9988
							1		5000	32	0.9989	0.9989	0.9989	0.9989	0.9988	0.9988	0.9988	0.9988
									50	32	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991
					lbfgs	l2	10	None	5000	32	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991	0.9991
									50	100	0.9984	0.9984	0.9984	0.9984	0.9983	0.9983	0.9983	0.9983
				Smote	saga	elasticnet	1	0.5	5000	358	0.999	0.999	0.999	0.999	0.9989	0.9989	0.9989	0.9989
									50	14	0.9991	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992
						None	1		5000	16	0.9991	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992
									50	15	0.9946	0.9946	0.9946	0.9946	0.9942	0.9942	0.9942	0.9942
							0.1		5000	16	0.9946	0.9946	0.9946	0.9946	0.9942	0.9942	0.9942	0.9942
									50	15	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984
							1		5000	14	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984	0.9984
									50	14	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
					lbfgs	l2	10	None	5000	16	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
									50	29	0.9974	0.9974	0.9974	0.9974	0.9997	0.9997	0.9997	0.9997
			Superclass	Rarest	saga	elasticnet	1	0.5	5000	109	0.9987	0.9987	0.9987	0.9987	0.9986	0.9986	0.9986	0.9986
									50	5	0.9162	0.9159	0.9162	0.916	0.939	0.9405	0.939	0.9394
						None	1		5000	5	0.9162	0.9159	0.9162	0.916	0.939	0.9406	0.939	0.9395
									50	6	0.9031	0.9031	0.9031	0.9029	0.9182	0.9202	0.9182	0.9186
							0.1		5000	7	0.9031	0.9031	0.9031	0.9029	0.9182	0.9202	0.9182	0.9186
									50	5	0.91	0.9097	0.91	0.9097	0.9286	0.9301	0.9286	0.929
							1		5000	6	0.91	0.9097	0.91	0.9097	0.9286	0.9301	0.9286	0.929
									50	5	0.9147	0.9143	0.9147	0.9144	0.9362	0.9377	0.9362	0.9367
					lbfgs	l2	10	None	5000	6	0.9147	0.9143	0.9147	0.9144	0.9362	0.9377	0.9362	0.9367
									50	13	0.9047	0.9046	0.9047	0.9044	0.9189	0.9208	0.9249	0.9193
				Smote	saga	elasticnet	1	0.5	5000	98	0.9118	0.9114	0.9118	0.9116	0.9317	0.9331	0.9317	0.9321
									50	1	0.9073	0.9066	0.9073	0.9069	0.9375	0.9385	0.9375	0.9377
						None	1		5000	2	0.9071	0.9065	0.9071	0.9068	0.9375	0.9386	0.9375	0.9377
									50	2	0.8685	0.8707	0.8685	0.8688	0.9043	0.9065	0.9043	0.9045
							0.1		5000	2	0.8685	0.8707	0.8685	0.8688	0.9043	0.9065	0.9043	0.9045
									50	3	0.8917	0.8918	0.8917	0.8914	0.9183	0.9197	0.9183	0.9185
							1		5000	2	0.8917	0.8918	0.8917	0.8914	0.9183	0.9197	0.9183	0.9185
									50	2	0.9	0.8996	0.9	0.8996	0.9277	0.9287	0.9277	0.9279
					lbfgs	l2	10	None	5000	2	0.9	0.8996	0.9	0.8996	0.9277	0.9287	0.9277	0.9279
									50	3	0.8908	0.8908	0.8908	0.8904	0.9163	0.9177	0.9163	0.9164
			Stars	Rarest	saga	elasticnet	1	0.5	5000	5	0.8935	0.8935	0.8935	0.8932	0.9197	0.9211	0.9197	0.9199
									50	11	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2188
						None	1		5000	12	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2188
									50	11	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2189
							0.1		5000	12	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2189
									50	11	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2189
							1		5000	10	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2189
									50	10	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2188
					lbfgs	l2	10	None	5000	11	0.3244	0.3001	0.3244	0.2966	0.2288	0.3009	0.2288	0.2188
									50	20	0.3244	0.3001	0.3244	0.2967	0.2288	0.3008	0.2288	0.2188
				Smote	saga	elasticnet	1	0.5	5000	22	0.3244	0.3001	0.3244	0.2967	0.2288	0.3008	0.2288	0.2188
									50	3	0.3195	0.2988	0.3195	0.2971	0.2296	0.2962	0.2296	0.2243
						None	1		5000	3	0.3195	0.2988	0.3195	0.2971	0.2296	0.2962	0.2296	0.2243
									50	3	0.3191	0.2979	0.3191	0.2965	0.2295	0.2962	0.2295	0.2242
							0.1		5000	3	0.3191	0.2979	0.3191	0.2965	0.2295	0.2962	0.2295	0.2242
									50	3	0.3199	0.2992	0.3199	0.2974	0.2295	0.296	0.2295	0.2242
							1		5000	3	0.3199	0.2992	0.3199	0.2974	0.2295	0.296	0.2295	0.2242
									50	3	0.3195	0.2989	0.3195	0.2971	0.2295	0.2961	0.2295	0.2242
					lbfgs	l2	10	None	5000	3	0.3195	0.2989	0.3195	0.2971	0.2295	0.2961	0.2295	0.2242
									50	3	0.3197	0.2989	0.3197	0.2972	0.2295	0.2959	0.2295	0.2241
Logistic	X0	1000	Galaxies	Rarest	saga	elasticnet	1	0.5	5000	4	0.3199	0.2991	0.3199	0.2974	0.2295	0.2958	0.2295	0.2241

Table 6 – Tuning table for support vector machine (tabular dataset)

								Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	C	Max Iterations	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
						10	54	0.8041	0.8726	0.8041	0.8084	0.8354	0.8721	0.8354	0.8313
					0.1	100	108	0.9151	0.922	0.9151	0.9149	0.9029	0.9119	0.9029	0.9031
						10	64	0.8041	0.8726	0.8041	0.8084	0.9029	0.9119	0.9029	0.9031
					1	100	122	0.7974	0.8398	0.7974	0.7842	0.722	0.8697	0.722	0.7309
						10	52	0.2979	0.4466	0.2979	0.1676	0.432	0.5731	0.432	0.3104
				Smote	10	100	114	0.7613	0.813	0.7613	0.7382	0.6698	0.8551	0.6698	0.6705
						10	22	0.7901	0.827	0.7901	0.7955	0.7966	0.8204	0.7966	0.7987
					0.1	100	40	0.9166	0.9216	0.9166	0.9166	0.9009	0.9118	0.9009	0.9022
						10	21	0.7517	0.8082	0.7517	0.7573	0.7623	0.7965	0.7623	0.7622
					1	100	38	0.9691	0.9702	0.9691	0.9691	0.9583	0.9633	0.9583	0.9592
						10	18	0.6497	0.6655	0.6497	0.6471	0.6885	0.6908	0.6885	0.6889
			Superclass	Rarest	10	100	31	0.975	0.9756	0.975	0.975	0.968	0.9717	0.968	0.9686
						10	6	0.5593	0.6003	0.5593	0.4949	0.4811	0.6187	0.4811	0.3952
					0.1	100	22	0.7277	0.8004	0.7277	0.7153	0.8	0.8297	0.8	0.788
						10	7	0.5682	0.6865	0.5682	0.5031	0.4842	0.7469	0.4842	0.3991
					1	100	19	0.7485	0.8076	0.7485	0.7151	0.8147	0.8476	0.8147	0.7931
						10	6	0.5651	0.5137	0.5651	0.495	0.4879	0.32113	0.4879	0.373
				Smote	10	100	17	0.7475	0.7786	0.7495	0.7508	0.8175	0.8282	0.8175	0.8163
						10	2	0.5994	0.7274	0.5994	0.5913	0.6841	0.7248	0.6841	0.6939
					0.1	100	5	0.8137	0.8248	0.8137	0.8131	0.8523	0.8643	0.8523	0.8518
						10	2	0.641	0.6967	0.641	0.6211	0.5844	0.7737	0.5844	0.578
					1	100	4	0.8773	0.8756	0.8773	0.8763	0.9106	0.9116	0.9106	0.9106
						10	2	0.6147	0.6998	0.6147	0.5516	0.5249	0.7018	0.5249	0.4256
			Stars	Rarest	10	100	3	0.8762	0.8966	0.8762	0.8683	0.9224	0.9353	0.9224	0.9175
						10	25	0.1031	0.1153	0.1031	0.0565	0.1376	0.1501	0.1376	0.1059
					0.1	100	180	0.0794	0.01424	0.0794	0.0476	0.1151	0.0958	0.1151	0.0794
						10	27	0.1036	0.0986	0.1036	0.0544	0.1347	0.146	0.1347	0.1
					1	100	182	0.098	0.2446	0.098	0.0692	0.1292	0.1348	0.1292	0.0936
						10	26	0.1379	0.1499	0.1349	0.079	0.1042	0.1052	0.1042	0.0509
				Smote	10	100	183	0.1042	0.1187	0.1042	0.0826	0.0963	0.1556	0.0963	0.0709
						10	6	0.1468	0.1808	0.1468	0.1213	0.1227	0.1783	0.1227	0.0973
					0.1	100	27	0.1846	0.1665	0.1846	0.1538	0.1489	0.1862	0.1489	0.1302
						10	6	0.138	0.1797	0.138	0.1077	0.1235	0.1784	0.1235	0.0991
					1	100	29	0.1762	0.1929	0.1763	0.1353	0.1303	0.1938	0.1303	0.107
						10	8	0.1395	0.146	0.1395	0.1096	0.1203	0.1904	0.1203	0.0907
SVM	X0	1000	Galaxies	Rarest	10	100	32	0.1445	0.1398	0.1445	0.1035	0.1302	0.1623	0.1302	0.0961

Table 7– Tuning table for decision tree (tabular dataset)

									Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	Criterion	Min Split	Max Depth	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
							2	31	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
						10	10	37	0.9998	0.9998	0.9998	0.9998	0.9994	0.9994	0.9994	0.9994
							2	32	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
						100	10	38	0.9996	0.9996	0.9996	0.9996	0.9993	0.9993	0.9993	0.9993
							2	32	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
					gini	1000	10	40	0.9993	0.9993	0.9993	0.9993	0.9991	0.9991	0.9991	0.9991
							2	31	0.992	0.9921	0.992	0.992	0.9951	0.9951	0.9951	0.9951
						10	10	37	0.9999	0.9999	0.9999	0.9999	0.9995	0.9995	0.9995	0.9995
							2	33	0.992	0.9921	0.992	0.992	0.9951	0.9951	0.9951	0.9951
						100	10	38	0.9997	0.9997	0.9997	0.9997	0.9995	0.9995	0.9995	0.9995
							2	34	0.992	0.9921	0.992	0.992	0.9951	0.9951	0.9951	0.9951
				Smote	entropy	1000	10	40	0.9991	0.9991	0.9991	0.9991	0.9988	0.9988	0.9988	0.9988
							2	14	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
						10	10	17	0.9997	0.9997	0.9997	0.9997	0.9993	0.9993	0.9993	0.9993
							2	15	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
						100	10	16	0.9994	0.9994	0.9994	0.9994	0.999	0.999	0.999	0.999
							2	15	0.9921	0.9922	0.9921	0.9921	0.9948	0.9948	0.9948	0.9948
					gini	1000	10	16	0.9971	0.9971	0.9971	0.9971	0.9956	0.9956	0.9956	0.9956
							2	14	0.992	0.9921	0.992	0.992	0.9941	0.9941	0.9941	0.9941
						10	10	16	0.9988	0.9988	0.9988	0.9988	0.9993	0.9993	0.9993	0.9993
							2	17	0.992	0.9921	0.992	0.992	0.9951	0.9951	0.9951	0.9951
						100	10	16	0.9993	0.9993	0.9993	0.9993	0.9989	0.9989	0.9989	0.9989
							2	16	0.992	0.9921	0.992	0.992	0.9951	0.9951	0.9951	0.9951
			Superclass	Rarest	entropy	1000	10	24	0.9979	0.9979	0.9979	0.9979	0.9981	0.9981	0.9981	0.9981
							2	5	0.5999	0.4588	0.5999	0.4966	0.8503	0.78	0.8503	0.8029
						10	10	6	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
							2	5	0.5999	0.4588	0.5999	0.4966	0.8503	0.78	0.8503	0.8029
						100	10	6	0.9976	0.9976	0.9976	0.9976	0.9976	0.9976	0.9976	0.9976
							2	6	0.5999	0.4588	0.5999	0.4966	0.8503	0.78	0.8503	0.8029
					gini	1000	10	7	0.9929	0.9929	0.9929	0.9929	0.9957	0.9957	0.9957	0.9957
							2	5	0.7798	0.6781	0.7798	0.7183	0.8749	0.8109	0.8749	0.8346
						10	10	6	0.9991	0.9991	0.9991	0.9991	0.9984	0.9984	0.9984	0.9984
							2	5	0.7798	0.6781	0.7798	0.7138	0.8749	0.8109	0.8749	0.8346
						100	10	6	0.9983	0.9983	0.9983	0.9983	0.9979	0.9979	0.9979	0.9979
							2	6	0.7798	0.6781	0.7798	0.7138	0.8749	0.8109	0.8749	0.8346
				Smote	entropy	1000	10	6	0.9924	0.9924	0.9924	0.9924	0.9953	0.9953	0.9953	0.9953
							2	2	0.6	0.4577	0.6	0.4961	0.8503	0.78	0.8503	0.8029
						10	10	3	0.9967	0.9967	0.9967	0.9967	0.9941	0.9941	0.9941	0.9941
							2	3	0.6	0.4577	0.6	0.4961	0.8503	0.78	0.8503	0.8029
						100	10	3	0.9907	0.9907	0.9907	0.9907	0.9924	0.9924	0.9924	0.9924
							2	2	0.6	0.4577	0.6	0.4961	0.8503	0.78	0.8503	0.8029
					gini	1000	10	2	0.9686	0.9708	0.9686	0.9684	0.9747	0.9777	0.9747	0.9749
							2	2	0.778	0.6751	0.778	0.7118	0.8749	0.8108	0.8749	0.8346
						10	10	2	0.9977	0.9977	0.9977	0.9977	0.9953	0.9953	0.9953	0.9953
							2	2	0.778	0.6751	0.778	0.7118	0.8749	0.8108	0.8749	0.8346
						100	10	2	0.9913	0.9913	0.9913	0.9913	0.9938	0.9938	0.9938	0.9938
							2	2	0.778	0.6751	0.778	0.7118	0.8749	0.8108	0.8749	0.8346
			Stars	Rarest	entropy	1000	10	2	0.9611	0.9637	0.9611	0.9605	0.9733	0.9766	0.9733	0.9734
							2	9	0.2198	0.0836	0.2198	0.1147	0.1284	0.0286	0.1284	0.0454
						10	10	11	0.3945	0.3821	0.3945	0.3799	0.2667	0.3308	0.2667	0.266
							2	9	0.2198	0.0836	0.2198	0.1147	0.1284	0.0286	0.1284	0.0454
						100	10	11	0.391	0.3779	0.391	0.376	0.2666	0.3312	0.2666	0.2659
							2	11	0.2198	0.0836	0.2198	0.1147	0.1284	0.0286	0.1284	0.0454
					gini	1000	10	11	0.3748	0.3616	0.3748	0.3568	0.2597	0.3334	0.2597	0.2568
							2	10	0.2138	0.0956	0.2138	0.1264	0.1456	0.1026	0.1456	0.0941
						10	10	12	0.3992	0.3744	0.3922	0.3761	0.2673	0.3271	0.2673	0.2669
							2	9	0.2138	0.0956	0.2138	0.1264	0.1456	0.1026	0.1456	0.0941
						100	10	13	0.3896	0.3719	0.3896	0.3735	0.2676	0.3277	0.2676	0.2669
							2	10	0.2138	0.0956	0.2138	0.1264	0.1456	0.1026	0.1456	0.0941
				Smote	entropy	1000	10	14	0.3732	0.3545	0.3732	0.3568	0.2645	0.3263	0.2645	0.2626
							2	3	0.223	0.0782	0.223	0.1122	0.1221	0.0298	0.1221	0.0453
						10	10	4	0.433	0.4311	0.433	0.4284	0.2559	0.2915	0.2559	0.26
							2	5	0.223	0.0782	0.223	0.1122	0.1221	0.0298	0.1221	0.0453
						100	10	4	0.3875	0.381	0.3875	0.3784	0.2576	0.2979	0.2576	0.2612
							2	3	0.223	0.0782	0.223	0.1122	0.1221	0.0298	0.1221	0.0453
					gini	1000	10	3	0.3084	0.2665	0.3084	0.2762	0.2158	0.2442	0.2158	0.2007
							2	3	0.2274	0.095	0.2274	0.1322	0.1812	0.0778	0.1812	0.1053
						10	10	5	0.4378	0.4361	0.4378	0.4332	0.2473	0.2952	0.2473	0.2557
							2	4	0.2274	0.095	0.2274	0.1322	0.1812	0.0778	0.1812	0.1053
						100	10	3	0.3827	0.3751	0.3827	0.3732	0.2528	0.3066	0.2528	0.2587
							2	4	0.2274	0.095	0.2274	0.1322	0.1812	0.0778	0.1812	0.1053
DT	X0	1000	Galaxies	Rarest	entropy	1000	10	4	0.315	0.303	0.315	0.2905	0.2245	0.2994	0.2245	0.2156

Table 8 – Tuning table for random forest (tabular dataset)

								Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	Bootstrap	# Estimators	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
						2	34	0.9999	0.9999	0.9999	0.9999	0.9994	0.9994	0.9994	0.9994
						10	70	0.9999	0.9999	0.9999	0.9999	0.9996	0.9996	0.9996	0.9996
					FALSE	50	189	0.9999	0.9999	0.9999	0.9999	0.9996	0.9996	0.9996	0.9996
						2	43	0.9996	0.9996	0.9996	0.9996	0.9994	0.9994	0.9994	0.9994
						10	58	0.9998	0.9998	0.9998	0.9998	0.9996	0.9996	0.9996	0.9996
				Smote	TRUE	50	158	0.9999	0.9999	0.9999	0.9999	0.9996	0.9996	0.9996	0.9996
						2	16	0.9999	0.9999	0.9999	0.9999	0.9993	0.9993	0.9993	0.9993
						10	23	0.9999	0.9999	0.9999	0.9999	0.9995	0.9995	0.9995	0.9995
					FALSE	50	58	0.9999	0.9999	0.9999	0.9999	0.9995	0.9995	0.9995	0.9995
						2	15	0.9996	0.9996	0.9996	0.9996	0.9993	0.9993	0.9993	0.9993
						10	20	0.9999	0.9999	0.9999	0.9999	0.9995	0.9995	0.9995	0.9995
			Superclass	Rarest	TRUE	50	44	0.9999	0.9999	0.9999	0.9999	0.9995	0.9995	0.9995	0.9995
						2	5	0.9985	0.9985	0.9985	0.9985	0.9971	0.9971	0.9971	0.9971
						10	8	0.998	0.998	0.998	0.998	0.9977	0.9977	0.9977	0.9977
					FALSE	50	21	0.9982	0.9982	0.9982	0.9982	0.9979	0.9979	0.9979	0.9979
						2	5	0.9959	0.9959	0.9959	0.9959	0.994	0.994	0.994	0.994
						10	7	0.9971	0.9971	0.9971	0.9971	0.9972	0.9972	0.9972	0.9972
				Smote	TRUE	50	16	0.9976	0.9976	0.9976	0.9976	0.9977	0.9977	0.9977	0.9977
						2	2	0.9983	0.9983	0.9983	0.9983	0.9874	0.9874	0.9874	0.9874
						10	2	0.9984	0.9984	0.9984	0.9984	0.9957	0.9957	0.9957	0.9957
					FALSE	50	2	0.9994	0.9994	0.9994	0.9994	0.9967	0.9967	0.9967	0.9967
						2	2	0.9924	0.9924	0.9924	0.9924	0.9872	0.9872	0.9872	0.9872
						10	2	0.999	0.999	0.999	0.999	0.996	0.996	0.996	0.996
			Stars	Rarest	TRUE	50	3	0.9989	0.9989	0.9989	0.9989	0.9962	0.9962	0.9962	0.9962
						2	12	0.3969	0.3797	0.3969	0.3795	0.272	0.3271	0.272	0.2707
						10	26	0.4106	0.3916	0.4106	0.3913	0.2775	0.3352	0.2775	0.2751
					FALSE	50	97	0.4147	0.3963	0.4147	0.3967	0.2889	0.337	0.2809	0.2806
						2	11	0.3891	0.3718	0.3891	0.3702	0.2634	0.3235	0.2634	0.2601
						10	15	0.4045	0.3866	0.4045	0.3872	0.2711	0.3298	0.2711	0.271
				Smote	TRUE	50	68	0.4143	0.3951	0.4134	0.3949	0.2799	0.3559	0.2799	0.2783
						2	2	0.4789	0.4761	0.4789	0.4706	0.2513	0.2944	0.2513	0.2487
						10	4	0.5532	0.5538	0.5532	0.5467	0.2647	0.3158	0.2647	0.2635
					FALSE	50	9	0.5777	0.578	0.5777	0.5719	0.2721	0.3287	0.2721	0.272
						2	3	0.4372	0.4295	0.4372	0.4307	0.2368	0.2812	0.2368	0.2405
						10	3	0.5383	0.5355	0.5383	0.5317	0.2627	0.3092	0.2627	0.2637
RF	X0	1000	Galaxies	Rarest	TRUE	50	6	0.564	0.5638	0.564	0.5576	0.2732	0.3254	0.2732	0.2739

Table 9 – Tuning table for feedforward neural network (tabular dataset)

								Train				Validation			
MODEL	X Dataset	Set Threshhold	Objects	Rebalancing Mode	Activations	Learning Rate	Run-Time (s)	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
						0.001	588	0.9995	0.9995	0.9995	0.9995	0.9994	0.9994	0.9994	0.9994
					relu	0.1	531	0.9996	0.9996	0.9996	0.9996	0.9994	0.9994	0.9994	0.9994
						0.001	587	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995
				Smote	tanh	0.1	528	0.9994	0.9994	0.9994	0.9994	0.9992	0.9992	0.9992	0.9992
						0.001	309	0.9996	0.9996	0.9996	0.9996	0.9995	0.9995	0.9995	0.9995
					relu	0.1	208	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995
						0.001	235	0.9993	0.9993	0.9993	0.9993	0.9993	0.9993	0.9993	0.9993
			Superclass	Rarest	tanh	0.1	232	0.9986	0.9986	0.9986	0.9986	0.9983	0.9983	0.9983	0.9983
						0.001	58	0.9722	0.9725	0.9722	0.9721	0.9759	0.9764	0.9759	0.9759
					relu	0.1	58	0.9832	0.9833	0.9832	0.9832	0.9832	0.9833	0.9832	0.9832
						0.001	59	0.9675	0.9681	0.9675	0.9674	0.9703	0.9713	0.9703	0.9704
				Smote	tanh	0.1	71	0.9376	0.9444	0.9376	0.9365	0.9526	0.9581	0.9526	0.9521
						0.001	17	0.7166	0.7349	0.7166	0.7067	0.7709	0.8151	0.7709	0.7762
					relu	0.1	12	0.8018	0.8001	0.8018	0.7988	0.8515	0.8593	0.8514	0.852
						0.001	16	0.881	0.8848	0.881	0.8816	0.9077	0.9116	0.9077	0.9076
			Stars	Rarest	tanh	0.1	16	0.9162	0.916	0.9162	0.9149	0.9279	0.931	0.9279	0.9281
						0.001	128	0.3769	0.3586	0.3769	0.3596	0.2782	0.3336	0.2782	0.2748
					relu	0.1	126	0.3703	0.3575	0.3703	0.3507	0.2804	0.3324	0.2804	0.2758
						0.001	182	0.374	0.3592	0.374	0.3589	0.3384	0.3384	0.2831	0.2802
				Smote	tanh	0.1	182	0.3538	0.3528	0.3538	0.3157	0.2504	0.3467	0.2504	0.2228
						0.001	31	0.3277	0.2938	0.3277	0.2799	0.23	0.2844	0.23	0.1986
					relu	0.1	21	0.2343	0.1726	0.2343	0.1792	0.148	0.1473	0.128	0.1048
						0.001	26	0.3324	0.2999	0.3324	0.2957	0.255	0.2938	0.255	0.234
FFNN	X0	1000	Galaxies	Rarest	tanh	0.1	22	0.2959	0.2735	0.2959	0.2524	0.2194	0.2693	0.2194	0.1888

Table 10 – Overview of models explored by classification task and dataset

	Classification Task			
Model	Superclass	Stars	Galaxies	Dataset
PCA	x	x	x	tabular
Linear Regression		x		tabular
K-means	x	x	x	tabular
KNN	x	x	x	tabular
Logistic Regression	x	x	x	tabular
SVM	x	x	x	tabular
Decision Tree	x	x	x	tabular
Random Forest	x	x	x	tabular
Feedforward NN	x	x	x	both
Res-Net			x	imagery
CNN			x	imagery
LSTM			x	imagery

Table 11 – Test subset metrics of tuned models for superclass

			Test				
Objects	Data Shape	MODEL	Run-Time (s)	Accuracy	Precision	Recall	F1
		KNN	156	0.9992	0.9992	0.9992	0.9992
		Logistic	15	0.9991	0.9991	0.9991	0.9991
		SVM	44	0.9568	0.9627	0.9568	0.9578
		DT	25	0.9992	0.9992	0.9992	0.9992
		RF	26	0.9994	0.9994	0.9994	**0.9994**
Superclass	Tabular	FFNN	252	0.9993	0.9993	0.9993	0.9993

Table 12 – Test subset metrics of tuned models for star subclass

			Test					
Objects	Data Shape	MODEL	Run-Time (s)	R^2	Accuracy	Precision	Recall	F1
		KNN	21		0.9819	0.9820	0.9819	0.9819
		Logistic	2		0.9346	0.9358	0.9346	0.9351
		SVM	3		0.9107	0.9164	0.9107	0.9078
		DT	2		0.9972	0.9972	0.9972	**0.9972**
		RF	2		0.9956	0.9956	0.9956	0.9956
Stars	Tabular	FFNN	62		0.9832	0.9836	0.9832	0.9833

Table 13 – Test subset metrics of tuned models for galaxy subclass

Objects	Data Shape	MODEL	Test Run-Time (s)	Accuracy	Precision	Recall	F1
		KNN	111	0.2693	0.3343	0.2693	0.2702
		Logistic	7	0.2242	0.3091	0.2242	0.2107
		SVM	31	0.1697	0.2031	0.17	0.1607
		DT	17	0.2639	0.3260	0.2639	0.2649
		RF	91	0.2844	0.3438	0.2844	0.2840
	Tabular	FFNN	144	0.3635	0.3565	0.3635	0.3412
		FFNN	3240	0.1548			
		ResNet50	34920	**0.5109**			
		CNN	7200	0.5034			
Galaxies	Imagery*	LSTM	5040	0.4397			

References

1 – Lilley, S. (1949). Pierre Simon Laplace (1749-1827). *Nature, 163, 468-469.*

2 - https://www.universetoday.com/164565/vera-rubin-will-generate-a-mind-boggling-amount-of-data/ (accessed 9 Dec. 2023)

3 - Lintott, C., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M., Nichol, R., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 389(3), 1179-1189.*

4 - https://www.sdss.org/science/ (accessed 9 Dec. 2023)

5 - https://www.sdss.org/instruments/ (accessed 9 Dec. 2023)

6 - Lupton, R., Zeljko Ivezic, Jim Gunn, Jill Knapp, Michael Strauss, & Naoki Yasuda (2002). The SDSS imaging pipelines. *Proceedings of SPIE - The International Society for Optical Engineering, 4836, 350–356.*

7 - https://data.galaxyzoo.org/ (accessed 9 Dec. 2023)

8 - Willett, K., Lintott, C., Bamford, S., Masters, K., Simmons, B., Casteels, K., Edmondson, E., Fortson, L., Kaviraj, S., Keel, W., Melvin, T., Nichol, R., Raddick, M., Schawinski, K., Simpson, R., Skibba, R., Smith, A., & Thomas, D. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 435(4), 2835–2860.*

9 - Hart, R., Bamford, S., Willett, K., Masters, K., Cardamone, C., Lintott, C., Mackay, R., Nichol, R., Rosslowe, C., Simmons, B., & Smethurst, R. (2016). Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society, 461(4), 3663–3682.*

10 – Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R., Raddick, M., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. (2011). Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies. *Monthly Notices of the Royal Astronomical Society, 410(1), 166-178.*

11 - Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. (2018). Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society, 476(3), 3661-3676.*

12 - Gharat, S., & Dandawate, Y. (2022). Galaxy classification: a deep learning approach for classifying Sloan Digital Sky Survey images. *Monthly Notices of the Royal Astronomical Society, 511(4), 5120–5124.*