Machine Learning Optimizes a Survey of Dark Energy

Steven Karst
Missouri University of Science and Technology

April 1, 2021

Table of Contents

Abstract

Measuring the history of cosmic expansion via the Baryon Acoustic Oscillation (BAO) scale from a three-dimensional galaxy map is a well established technique to probe the nature of dark energy. In fact, a forthcoming galaxy redshift survey, the Subaru Prime Focus Spectrograph (PFS), is designed mainly for this purpose. An essential optimization problem in such galaxy redshift surveys is the target selection. Namely, it is not clear how we should select our targets to maximize the number of galaxies which provide successful redshift measurement in a desired cosmic epoch, while avoiding other galaxies. Taking PFS as an example, we apply a modern machine learning algorithm to the target selection problem. In this project we analyze how well machine learning could optimize the PFS survey target selection compared to more conventional methods, and show that our new approach could play a crucial role in understanding dark energy.

Introduction

A current mystery in the field of cosmology is the nature of dark energy, a mysterious energy component that is counteracting gravity and causing the expansion of the Universe to accelerate over time. A common and established technique to measure the nature of dark energy is to measure the distance to galaxy populations using the Baryon Acoustic Oscillation (BAO) scale as a standard ruler. This technique has been successfully used in the Sloan Digital Sky Survey (eBOSS Collaboration, 2020) and is planned to be used several upcoming galaxy surveys such as the Hobby-Eberly Telescope Dark Energy Experiment (Hill et al., 2005), the Dark Energy Spectroscopic Instrument or DESI (Karim et al., 2020) and the Subaru Prime Focus Spectrograph or PFS (Takada et al., 2014). Each of these surveys specifically attempt to measure the redshifts of star-forming galaxies through strong emission lines.

An important step in each of these surveys is the target selection. To efficiently measure emission lines, galaxies that are likely to have sufficiently strong emission lines in a certain redshift range must be pre-selected from photometric images by measuring along certain photometric bands. In DESI, for example, galaxy flux measurements within the photometric bands g, r and i are analyzed within an imaging dataset to target [OII] emitters with z_red values - or fractional changes in photon wavelength - less than 1.6 (Karim et al., 2020). PFS meanwhile has the advantage of having a near-infrared camera that can analyze the fluxes within the photometric bands g, r, i, z and y - with galaxy specific measurements denoted as *g, r, i, z* and *y* - to target [OII] emitters with z_red values between 0.6 and 2.4 in Hyper-Suprime Cam (Takada et al., 2014). The locations of these bands are marked in Figure 1.

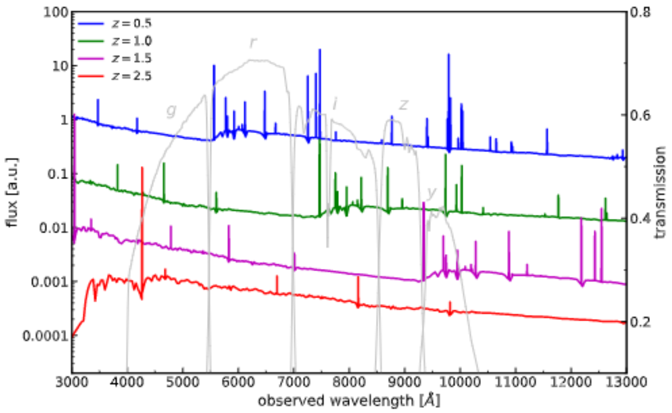

Figure 1: Modeled galaxy spectra at different redshifts in EL-COSMOS (Saito et al., 2020). The main target is the [OII] 3726-3729 line at a wavelength of (1+z)*3726 Angstroms, which is shown in the colored lines.

In addition, the PFS cosmology program is designed to have two visits per 1.098 degrees squared field-of-view (FoV) with a 15-minute exposure time, allowing 4788 fibers to be used per FoV. Because of this limited number of available spectrographic fibers, the regions of space analyzed by PFS must be as clear from non-target galaxies as possible, while still containing several target galaxies (ideally 100% of the target galaxies in the catalog). The distribution of target and non-target galaxies in EL-COSMOS, a mock galaxy catalog (Saito et al., 2020), can be seen in Figure 2:

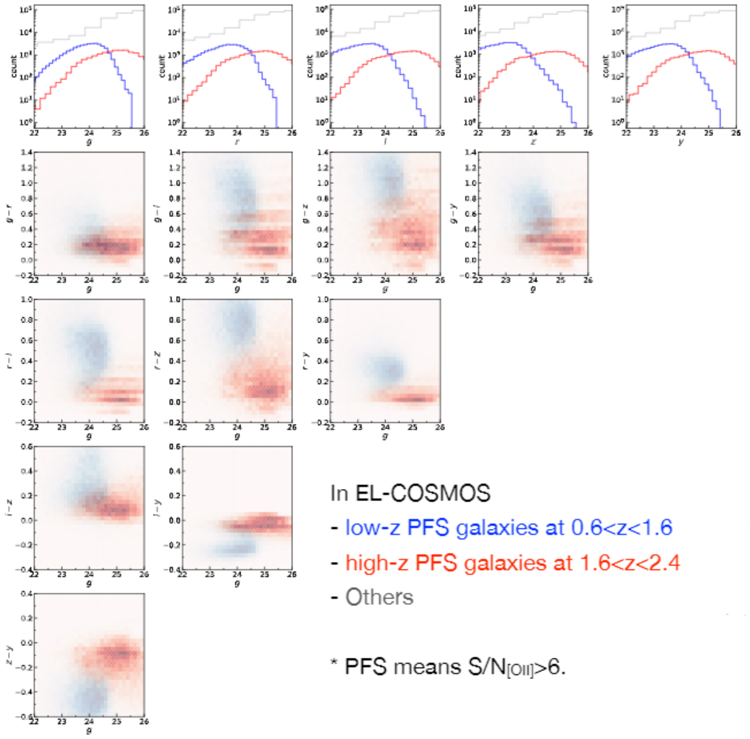

Figure 2: A histogram of galaxies in EL-COSMOS along several combinations of photometric bands.

Though Figure 2 shows the clustering of the data in parts, the high dimensionality of this data means that it is difficult to visualize fully in a single representation. This means that efficiently drawing a decision region within *g, r, i, z* and *y* is a major challenge. In the field of data science, such high-dimensional datasets are often explored with machine learning algorithms such as multilayer perceptrons and decision trees, which can be trained in a computer program to detect and conform to patterns even in large datasets. Recently, these machine learning algorithms have seen use in astrophysical and cosmological research, such as signal processing in gravitational wave analysis.

With this in mind, we propose using machine learning algorithms to complete an exploratory analysis of a mock catalog and learn patterns that can distinguish between target and non-target galaxies. In this paper, several machine learning classifiers of varying parameters are used to separate a mock dataset into non-target galaxies, Low-Z (0.6<z_red<1.6) target galaxies, and High-Z (1.6<z_red<2.4) target galaxies, with the goal of maximizing the number of collected galaxies and the percentage of target galaxies within a sample, with a particular emphasis on High-Z galaxies to make full use of the capabilities of PFS compared to experiments that cannot read High-Z galaxies.

Hypothesis

If machine learning classifiers are trained on a mock catalog of PFS target and non-target galaxies and used to estimate target regions, then the classifiers will outperform approaches that do not utilize machine learning in terms of collecting a large sample of target galaxies with a large amount of High-Z galaxies.

Methodology

At first, the EL-COSMOS Catalog (Saito et al., 2020) was used to train and test several machine learning classifiers on *g, r,* and *i*. Each galaxy within the catalog was labeled as 0 if it was not a PFS target galaxy, or 1 if the galaxy was a PFS target. The catalog was also randomly separated into training and testing datasets such that 70% of galaxies within the catalog were used to build algorithm decision boundaries, while the remaining 30% were used to test the algorithms and ensure the classifiers could accurately label data that was unseen during training. Explanations of these algorithms can be found in the nomenclature section of this paper.

Random forests were then selected for further testing due to the relative ease of training them and their competitive accuracy. Several random forests were trained on different combinations of *g*, *r, i, z* and *y* values. Some algorithms were also trained on a mutation of the original labeling procedure where class 1 (target) galaxies with a z_red value above 1.6 were instead labeled 2, marking them as High-Z galaxies. This mutation was called the triple-class label. Some algorithms also used a version of the catalog where missing values within the catalog were imputed via a median-centered simple imputer.

After this optimization step, a random forest trained on *g, g-r, r-i, i-z* and *z-y* with missing values imputed in became the baseline forest for further testing. The importance of a feature within the random forest was defined as the percentage of nodes within the forest that use the given feature to draw a boundary. The importance of each feature within the baseline forest was collected and compared to the importances of a forest trained on the imputed catalog and only on *g, g-r* and *r-i* to determine if the dim z and y bands gave enough information to be useful in target classification.

The baseline forest was then used to test several visualization techniques that could give qualitative information about how the baseline forest defined its decision regions, such as a Voronoi Tessellation plot of a PCA projection (Migut et al., 2015). Because the input dataset was 5-dimensional, these techniques were either animated or projected into a lower feature space. Using these results, outliers were detected and the general range of the dataset was analyzed.

Finally, the decision boundaries of the baseline forest were used to create a set of 5-cubes defining regions with high amounts of target galaxies and low amounts of non-target galaxies. The High-Z galaxies within the training dataset were used to generate a 5D histogram. The peak of this histogram was then extracted and separated further into 5-cubes. The centers of these 5-cubes were then sent to the baseline forest for classification. Any 5-cubes that had their center labelled as a target were then considered to be target regions that PFS should aim for during its galaxy collection. This process was then repeated with the next peak of the histogram, and repeated more until at least 5700 galaxies in the catalog fell within the boundaries. The process was then repeated using the Low-Z target histogram.

Results

Table 1 compares the accuracy of several machine learning classifiers trained and tested on the galaxy catalog. Each classifier was trained in a dual-class setting and only on *g, r* and *i*:

Classifier	Hyperparameters	Train Acc.	Test Acc.
Random Forest	25 decision trees of max depth 25	99.53%	95.54%
KNN	Minkowski metric with p=2, 5 neighbors considered	96.34%	94.95%
AdaBoost	0.1 learning rate 500 decision tree classifiers of depth 1	93.68%	93.60%
Bagging	25 decision trees of max depth 25	99.49%	95.56%

Table 1: Results of different classifiers trained on *g, r* and *i*.

The random forest was then retrained multiple times with different parameters, as shown in Table 2:

Dimensions	Imputes?	Classes	Train Acc.	Test Acc.
g, g-r, r-i	No	Dual-Class	99.09%	95.16%
g, r, i, z, y	No	Dual-Class	99.73%	95.98%
g, r, i, z, y	No	Triple-Class	97.57%	94.60%
g, r, i, z, y	Yes	Dual-Class	99.12%	95.70%
g, g-r, r-i, i-z, z-y	No	Dual-Class	99.55%	97.46%
g, g-r, r-i, i-z, z-y	Yes	Triple-Class	99.97%	97.64%

Table 2: Results of different random forests trained on the EL-COSMOS catalog. Every forest contained 25 decision trees.

Table 3 shows the confusion matrix of the baseline forest of predictions using the test dataset:

Baseline Forest Confusion Matrix on Test Data		True Class		
		Non-Target	Low-Z Target	High-Z Target
Predicted Class	Non-Target	137602	801	479
	Low-Z Target	1193	10087	25
	High-Z Target	1191	43	4101

Table 3: A confusion matrix of the baseline forest for the test dataset.

Figure 3 compares the importances of the baseline forest and a forest trained on g, g-r, and r-i:

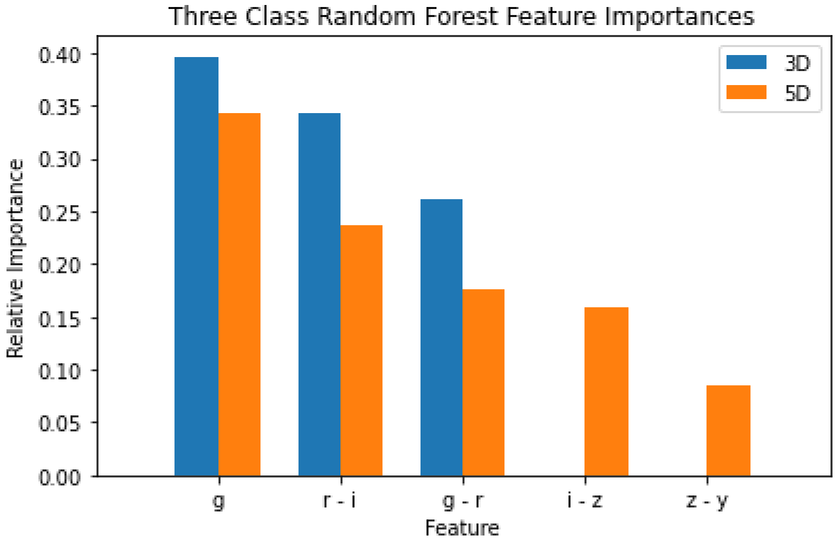

Figure 3: A comparison of the feature importances, or percentage of tree nodes splitting based on a feature, between a 3D baseline forest and a 5D baseline forest.

Figure 4 compares the false positives and true positives found by a baseline forest trained on *g, r, i, z* and *y* on the left to the false and true positives found by a baseline forest trained on *g, g-r, r-i, i-z* and *z-y* on the right:

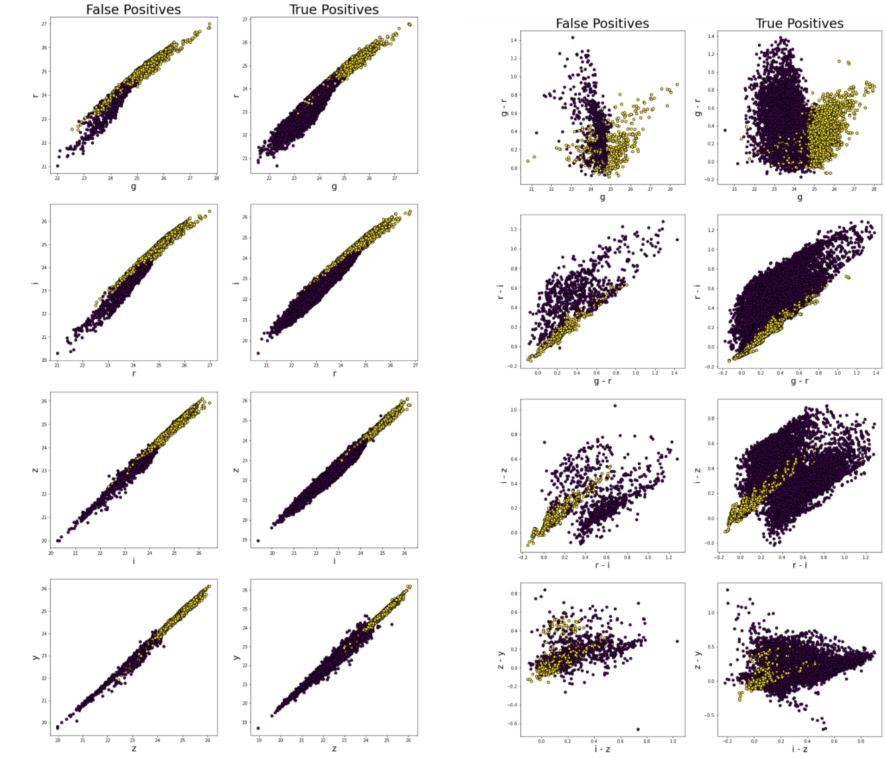

Figure 4: A comparison of the distribution of false and true positives between a baseline forest trained on individual photometric bands on the left and a baseline forest trained on differences between photometric bands on the right. Low-Z predictions are marked in purple and High-Z predictions are marked in yellow.

Figure 5 shows the results of the baseline forest projected into two dimensions using PCA (principal component analysis) with decision boundaries estimated using a Voronoi Tessellation plot (Migut et al., 2015):

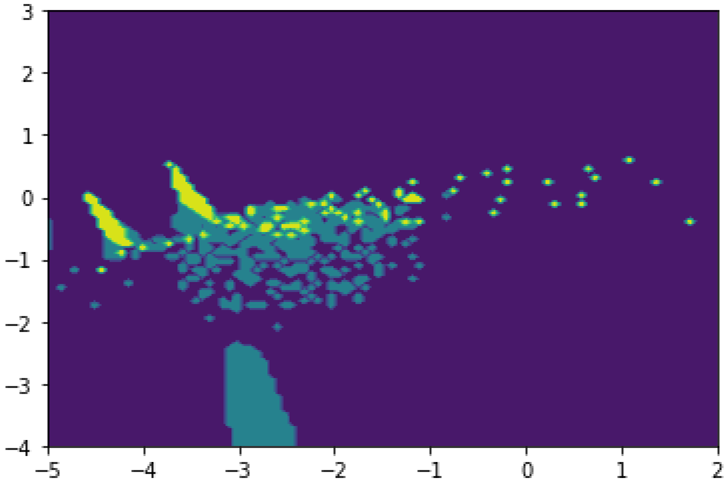

Figure 5: A Voronoi Tessellation plot (Migut et al., 2015) that estimates the decision boundary of the baseline forest using two principal components found in PCA. Low-Z target regions are marked in blue, and High-Z target regions are marked in yellow.

Figure 6 compares the distribution of z_red values for the collected galaxies falling within the 5-cube estimated regions of the machine learning classifiers to other methods used in the past. The red line represents 5-cubes drawn using the entire target galaxy histogram and the purple line represents the 5-cubes with half of the collected galaxies collected only from the High-Z target histogram:

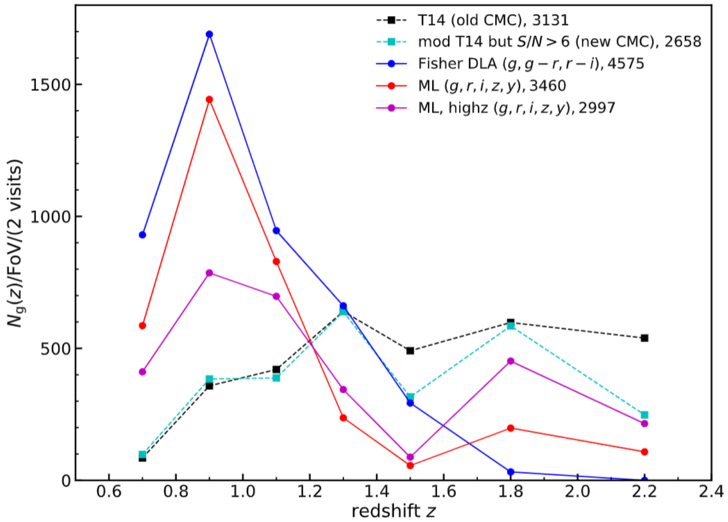

Figure 6: A comparison of several galaxy distributions collected from the EL-COSMOS catalog (Saito et al., 2020) using different methods of drawing decision regions.

Discussion

As shown in Table 1, the two most competitive classifiers were a random forest and a bagging ensemble. Adaboost had both the lowest training accuracy and test accuracy, while KNN had the next lowest accuracy on both training and testing. It should be noted that the catalog is an imbalanced dataset, with non-target galaxies representing about 90% of the entire dataset; therefore, even a percentage point of difference between the accuracies of two classifiers is significant. Even with this in mind, bagging and random forest classifiers performed similarly, with bagging performing slightly better on testing. However, random forests were still chosen as the main classifier for this paper due to their simplicity, speed and readability compared to bagging.

Of the random forests compared in Table 2, it was determined that a forest trained on the imputed catalog and on *g, g-r, r-i, i-z* and *z-y* was most optimal, so this forest was used as the baseline classifier for the rest of the research. Interestingly, while training on only *g, g-r* and *r-i* produced a classifier with lower accuracy on dual-class data compared to a classifier trained on all five bands individually, this accuracy decreased significantly in a triple-class situation. However, training on *g, g-r, r-i, i-z* and *z-y* produced significantly more accurate predictions than any other classifier, and had further increases in accuracy when trained on the imputed catalog and in a triple-class setting. Therefore, the differences between the fluxes along the photometric bands are needed to accurately determine the location of High-Z galaxies.

This idea is shown further in Figure 4. Along combinations of individual dimensions, Low-Z and High-Z true positives always followed a linear pattern and had significant overlap. False positives also followed this pattern and were close together, indicating that the catalog itself likely has this linear pattern. However, when differences between dimensions were taken into account, both Low-Z and High-Z true positives had clearly separable distributions, while the sparsity of the false positives indicates that those mistakes are likely noise that signs that the entire dataset follows the pattern picked up by the random forest.

Figure 3 highlights that the dim z and y photometric bands contain enough information to justify analyzing them during galaxy selection. As shown in the graph, cuts along *i-z* and *z-y* were present in about 25% of the nodes within the baseline forest. The relative importances of *g, g-r* and *r-i* were also all smaller in the baseline forest compared to the forest trained on *g, g-r* and *r-i*. Furthermore, Table 2 shows that the baseline forest was nearly 2.5% more accurate than a forest trained only on *g, g-r* and *r-i*. Therefore, *i-z* and *z-y* were kept during the decision region estimation. The baseline forest confusion matrix shown in Table 3 also indicates that the baseline forest could accurately identify about 90% of the target galaxies within the catalog at about 90% confidence showing that, on its own, the baseline forest is very capable of sorting the galaxy dataset and collecting patterns within it.

As an abstract way of showing the decision boundaries, Figure 5 indicates that, while Low-Z galaxies are usually separable from non-target galaxies, High-Z galaxies mostly fall within Low-Z boundaries, with some sparsely spread out within the non-target galaxies. This helps to explain why High-Z galaxies were more difficult to predict from the training data, as their distribution was more spread out and significantly intersected with both Low-Z and non-target regions.

Figure 6 indicates that the machine learning approach presented within this paper is competitive with other methods used in the past. Before this approach, most target selection methods performed on the catalog had a success rate (defined as the percentage of galaxies collected by PFS in the decision regions that are target galaxies) of 0.5 or 0.6, while the machine learning approach had a success rate of 0.66, which shows a modest improvement. However, less galaxies with a relatively high redshift were collected with the cube-drawing approach compared to previous methods, though forcing some cubes to be based around the High-Z target histogram helped somewhat in this regard.

The machine learning approach is not without limitations. For example, the current approach of estimating the decision boundaries of the baseline forest from 5-cubes sacrifices much of the information contained within the classifier, and takes a relatively long time to complete. The cubes are currently drawn at a resolution of 0.2 along each individual band, but reducing this resolution to make the estimation more accurate increases the complexity of the algorithm to an unreasonable degree. If this approach is used in the future, work needs to be done to get as close as possible to calculating the direct decision boundaries from the classifier.

Furthermore, the 5-cubes are not necessarily adjacent to each other, so the decision boundary found with the current approach is disjointed. The success rate is also calculated from the entire dataset rather than just the testing dataset. This means that the 5-cubes rely heavily on the simulated catalog being correct, which is not a guarantee. Outside analysis of the cubes indicates that the proportion of test galaxies falling within the boundaries is about equal to that of the training galaxies, so the algorithm does have working protections against overfitting to the data, but work should be done on adjusting the algorithm to create one continuous decision region.

In conclusion, the machine learning approach shown in this paper is a competitive way to draw decision boundaries during target selection, so continuing work on implementing the algorithms is worthwhile. While estimating the decision boundaries from the classifiers still needs to be worked on to surpass other methods in collecting High-Z target galaxies, the base classifiers have been shown to accurately distinguish between target and non-target galaxies, while also detecting significant patterns in the dataset. In the future, the estimation of the decision boundaries from machine learning algorithms should be more based on direct extraction from the algorithm rather than simple estimation, and new algorithms such as neural networks may also be important to study for PFS target selection. Therefore, while the hypothesis was only partially correct in that decision boundaries based on machine learning algorithms had a higher success rate than conventional methods but collected slightly less High-Z target galaxies, this could change with more experimentation.

Nomenclature

Baryon Acoustic Oscillation (BAO) Scale: A method of measuring cosmological distances using the propagation of sound waves in primordial plasma present in the early universe, which allows cosmological expansion to be accounted for in measurements.

[OII] Emission Line: The strength of light energy originating from singly ionized oxygen at different wavelengths.

Photometric Band: An interval of wavelengths and maximum flux values where the flux of a galaxy's light emissions can be measured and analyzed.

Feature: A dimension that a machine learning algorithm is trained on to predict class labels or measurements of a dependent variable.

KNN: Abbreviation for K-Nearest Neighbor, a machine learning algorithm that classifies a given sample by comparing it to the *k* nearest samples in the training dataset, where *k* represents a positive integer.

Bagging: An ensemble of classifiers that are each trained on random samples of the training dataset taken with replacement. The ensemble classifies samples using a majority vote among its parts, and often has simpler decision boundaries than what would be expected from an individual classifier.

Decision Tree: A machine learning algorithm consisting of several interconnected nodes representing one-dimensional if-else statements designed to reduce data impurity as much as possible.

Random Forest: A special case of bagging using decision trees that also take random samples of the features of the training dataset, while keeping the underlying decision tree structure intact.

AdaBoost: An ensemble of extremely weak classifiers trained in series such that the mistakes of earlier classifiers punish later classifiers during training, often resulting in more complex decision boundaries than in the base classifier.

5-Cube: A hypercube defined in a 5-dimensional space such that the shape contains 80 equally sized edges.

Acknowledgments

The author would like to sincerely thank the following people and groups for their contributions to this project:

Dr. Shun Saito, the research advisor on this project;

The *Missouri S&T National Merit Semifinalist Scholarship Package,* the source of funding for this project;

The *PFS Target Selection Discussion Group*, whose members provided direct feedback on the procedure and results of this project;

Scikit-learn, the Python library used for training the machine learning algorithms in this project (Pedregosa et al., 2011);

Numpy, the Python library used for efficient array manipulation in this project under the Creative Commons 4 license (Van Der Walt, 2011);

Matplotlib, the Python library used to generate the graphs in this project (Hunter, 2007);

JupyterLab and *Google Colab*, the IDEs used for programming in this project (Granger & Grout, 2016);

Anaconda, the Python environment manager used to download the necessary Python libraries;

And *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition*, the textbook used as a reference for implementing machine learning in this project (Raschka and Mirjalili, 2016).

References

eBoss Collaboration: Alam, S., Aubert, M., Avila, S., Balland, C., Bautista, J. E., Bershady, M. A., ... & Zheng, Z. (2020). The completed SDSS-IV extended baryon oscillation spectroscopic survey: cosmological implications from two decades of spectroscopic surveys at the apache point observatory. *arXiv preprint arXiv:2007.08991*.

Granger, B., & Grout, J. (2016). JupyterLab: Building Blocks for Interactive Computing. *Slides of presentation made at SciPy*.

Hill, G., Gebhardt, K., MacQueen, P. J., & Komatsu, E. (2005). Hobby-Eberly Telescope Dark Energy Experiment (HETDEX). *Proceedings of" Probing the Dark Universe with Subaru and Gemini*, 22-1.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *IEEE Annals of the History of Computing*, *9*(03), 90-95.

Karim, T., Lee, J. H., Eisenstein, D. J., Burtin, E., Moustakas, J., Raichoor, A., & Yèche, C. (2020). Validation of emission-line galaxies target selection algorithms for the Dark Energy Spectroscopic Instrument using the MMT Binospec. *Monthly Notices of the Royal Astronomical Society*, *497*(4), 4587-4601.

Migut, M. A., Worring, M., & Veenman, C. J. (2015). Visualizing multi-dimensional decision boundaries in 2D. *Data Mining and Knowledge Discovery*, *29*(1), 273-295.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Birmingham, United Kingdom: Packt.

Saito, S., de la Torre, S., Ilbert, O., Dubois, C., Yabe, K., & Coupon, J. (2020). The synthetic Emission Line COSMOS catalogue: Hα and [O ii] galaxy luminosity functions and counts at 0.3< z< 2.5. *Monthly Notices of the Royal Astronomical Society*, *494*(1), 199-217.

Takada, M., Ellis, R. S., Chiba, M., Greene, J. E., Aihara, H., Arimoto, N., ... & Wyse, R. (2014). Extragalactic science, cosmology, and Galactic archaeology with the Subaru Prime Focus Spectrograph. *Publications of the Astronomical Society of Japan*, *66*(1).

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering*, *13*(2), 22-30.