Performance Evaluation of Random Forest and XGBoost for Multiclass Classification of Astronomical Objects: SDSS DR18 Analysis

Vignesh Nadar[1], Zainab Shaikh[2]

[1] M.Sc.IT, SIES (Nerul) College of Arts, Science and Commerce

[2] M.Sc.IT, SIES (Nerul) College of Arts, Science and Commerce

---

Abstract:

The vast dataset provided by the Sloan Digital Sky Survey (SDSS) presents an invaluable resource for the classification of celestial objects such as stars, galaxies, and quasars. This study aims to evaluate the effectiveness of two powerful machine learning algorithms, XGBoost and Random Forest Classifier, in handling the multiclass classification tasks within this extensive dataset. Leveraging the advanced capabilities of SDSS Data Release 18 (DR18), which includes comprehensive photometric and spectroscopic data, we meticulously implemented and compared these algorithms.

Our findings corroborate prior research, establishing that both XGBoost and Random Forest Classifier demonstrate robust performance in classifying astronomical objects. Despite thorough fine-tuning efforts, no significant improvements in accuracy were observed, with XGBoost even showing a marginal decline, indicating that these models were already optimized for the task. This study underscores the reliability and efficiency of these algorithms in analyzing SDSS data, emphasizing their pivotal role in astronomical research. These results contribute to the broader understanding of machine learning applications in astronomy and suggest that while XGBoost and Random Forest Classifier are highly effective, future studies might explore additional features or hybrid models to enhance classification accuracy further.

In summary, this study reaffirms the utility of XGBoost and Random Forest Classifier in the classification of celestial objects within the SDSS dataset, highlighting their robustness and reliability. It provides a comprehensive analysis that can inform future astronomical data classification efforts, paving the way for continued advancements in the field.

Keywords:

Astronomy, Supervised learning, Classification, SDSS, Sky survey, EDA, Feature selection, Random forest classifier, XGBoost, Optimization/Hypertuning, Multiclass classification.

1. Introduction

The Sloan Digital Sky Survey (SDSS), launched in 2000, stands as a cornerstone in our exploration of the cosmos. Nestled within the Apache Point Observatory in New Mexico, USA, SDSS employs cutting-edge technology, including a 120-megapixel camera and advanced spectrographs, to meticulously map the celestial landscape. With its reach spanning more than a quarter of the sky, SDSS has cataloged an impressive array of celestial objects, ranging from stars and galaxies to elusive quasars.

Over the years, SDSS has undertaken multiple phases, each with its own distinct objectives and areas of focus. These phases include the original SDSS-I (2000-2005), which laid the groundwork for subsequent observations, and SDSS-II (2005-2008), which expanded the survey's scope to include studies of the Milky Way's structure and the search for quasars.

Subsequent phases, such as SDSS-III (2008-2014) and SDSS-IV (2014-present), have continued to push the boundaries of astronomical research, delving into topics ranging from the nature of dark energy and the large-scale structure of the universe to the dynamics of galaxy evolution and the identification of rare celestial phenomena.

However, the sheer volume of data amassed by SDSS renders manual classification impractical. As a result, automated methods, particularly machine learning techniques have become indispensable. While unsupervised methods like k-means clustering have been employed successfully, this paper advocates for the use of supervised learning algorithms. Specifically, it proposes to leverage XGBoost, a powerful gradient boosting algorithm, alongside Random Forest (RF).

XGBoost, renowned for its ability to handle complex datasets and deliver superior predictive performance, is poised to offer valuable insights into the classification of celestial objects. Coupled with RF, an ensemble method renowned for its robustness and versatility, this study aims to provide a comprehensive comparison of these algorithms' efficacy. By focusing on a subset of 100,000 data points from the SDSS database, we endeavor to unravel the unique strengths and weaknesses of XGBoost and RF in the domain of celestial object classification.

2. Literature Review

The Sloan Digital Sky Survey (SDSS) has been a pivotal resource in astronomical research since its inception in 2000. Its comprehensive sky surveys have provided invaluable data for classifying celestial objects. The SDSS phases, ranging from SDSS-I to the current SDSS-V, have progressively enhanced the scope and depth of astronomical observations, making substantial contributions to our understanding of the cosmos.

The application of machine learning (ML) techniques in astronomy has grown significantly, driven by the need to manage and analyze large datasets like those from SDSS. Traditional methods of classification have been complemented and, in many cases, supplanted by supervised learning algorithms due to their accuracy and efficiency in handling complex data.

Random Forest (RF) and XGBoost are two of the most prominent algorithms used for classification tasks in astronomy. Random Forest, an ensemble method based on decision trees, is renowned for its robustness and capability to handle noisy data. It aggregates the predictions of multiple trees to improve generalization and reduce overfitting.

XGBoost, or eXtreme Gradient Boosting, is a gradient boosting algorithm that has gained popularity for its high performance and efficiency. It constructs an ensemble of decision trees iteratively, optimizing a specific loss function and using gradient descent. XGBoost is known for its speed and accuracy, especially in large datasets.

Several studies have highlighted the effectiveness of RF and XGBoost in classifying astronomical objects. For instance, Talla et al. demonstrated the efficacy of Random Forest in multiclass classification, emphasizing its predictive power and stability. Similarly, Mohbey et al. showcased XGBoost's superior performance in credit card fraud detection, underlining its versatility and robustness across different domains.

In the context of astronomical data, Rafid Bendimerad's work on supervised classification using SDSS measurements underscored the significance of using both photometric and spectroscopic data to enhance classification accuracy.

The current study builds upon these foundational works by comparing the performance of RF and XGBoost specifically for classifying stars, galaxies, and quasars using the SDSS DR18 dataset. The investigation revealed that both algorithms performed exceptionally well, with no significant improvement in accuracy observed upon hyperparameter tuning. This suggests that the models were already optimized for the task at hand.

3. Methodology

3.1 SDSS Data

The SDSS stands as a cornerstone dataset, providing a wealth of information for celestial object classification. It encompasses both photometric and spectroscopic data, capturing the electromagnetic emissions of millions of celestial sources. The photometric data span multiple wavelength bands, including ultraviolet, optical, and infrared, enabling a comprehensive exploration of the diverse properties of stars, galaxies, and quasars. Additionally, the spectroscopic observations from SDSS offer detailed spectral lines and redshift information crucial for understanding the chemical composition and cosmological distances of celestial objects.

In this paper we will be working with the Data Release 18 (DR18) dataset. DR18 is the first data release for the fifth phase of the Sloan Digital Sky Survey (SDSS-V).

DR18 includes the following:

- Targeting catalogs prepared for the Black Hole Mapper and Milky Way Mapper science programs, as well as for the open fiber programs
- Black Hole Mapper eFEDs spectra, part of the SPIDERS science program
- A Value Added Catalog of redshifts and classifications for all DR18 eFEDS targets

This data can be easily collected from the SkyServer tool. It's a very intuitive tool for exploration of different celestial bodies in different releases and collection of these data using SQL. The dataset once queried using sql can be downloaded in any required formats such as FITS or csv. In this paper we have used SQL queries in order to access and download 100,000 rows of SDSS data in csv format. This dataset has 18 different features namely objid, ra, dec, u, g, r, i, z, run, rerun, camcol, field, specobjid, class, redshift, plate, mjd and fiberid.

Once the dataset has been downloaded now we can begin our steps towards understanding the data and using it in the most efficient way and visualize it whenever necessary, in order to understand the different galactic objects such as stars, galaxies and quasars.

3.2 Univariate Analysis

Univariate analysis involves the examination of a single variable at a time, aiming to understand its distribution, characteristics, and patterns without considering its relationship with other variables. Through descriptive statistics, data visualization, and frequency analysis, univariate analysis provides valuable insights into the central tendency, dispersion, and shape of the variable's distribution. This foundational step in exploratory data analysis helps identify outliers, detect patterns, and guide further investigation or modeling efforts in data analysis and research.

Redshift is a fundamental concept in astronomy and cosmology that describes

how light from distant celestial objects, such as galaxies and quasars, is shifted towards longer wavelengths as they move away from the observer. To start the univariate analysis we will plot histograms for the 'redshift' feature column for each class. This will tell us how the redshift values are distributed over their range.

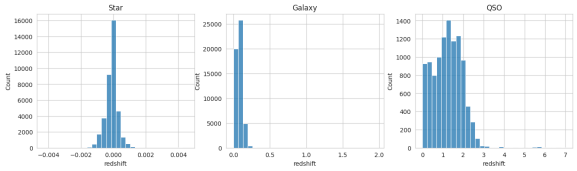

Fig. 1. Plots showing how redshift values are distributed for stars, galaxies and quasars.

The above plots show how different celestial bodies gather around specific redshift values. This indicates that each object be it star, galaxy or a quasar, has its own characteristics redshift value.

- The stars seem to have a redshift value closer to 0. The distribution looks like a thin bell curve.
- The galaxies seem to have redshift values between 0 and ~0.3. This distribution is a bit more uniform than the stars.
- Quasars have the most uniform distribution of redshift values. They seem to range from 0 to 3, and we can see a few outliers after that.

This uniqueness of the redshift feature ensures that it can be very helpful in the classification process.

3.3 Multivariate Analysis

Multivariate analysis is a statistical method used to analyze datasets that involve multiple variables simultaneously. Unlike univariate analysis, which focuses on examining the characteristics of individual variables in isolation, multivariate analysis considers the relationships between multiple variables and aims to understand the complex interactions and patterns within the data.

In multivariate analysis, the goal is typically to explore how different variables are related to each other, identify underlying structures or patterns in the data, and make predictions or inferences based on these relationships.

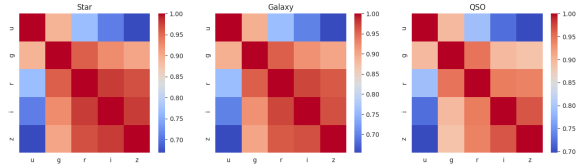

Fig. 2. These plots show how the different wavelength bands (u, g, r, i, z) are correlated with each other.

The correlation plots look very similar to each other. This could be because the majority of them may be constituted of similar components, say hydrogen, carbon, oxygen or any such elements which are known to be the building blocks of anything in this universe.

One interesting part is that the u band seems to be less correlated with other bands, whereas other bands are well correlated with each other.

3.4 Representation of the data in celestial coordinates

The positions of observed celestial objects are determined by their right ascension and declination coordinates,

forming a crucial reference framework for locating objects in the sky. These coordinates provide a standardized way to assign precise locations to each observed object, facilitating efficient cataloging and analysis of astronomical data. The data collected by the SDSS primarily comes from the northern celestial hemisphere due to the geographic location of the SDSS telescope at the Apache Point Observatory in New Mexico, USA. The telescope's position in the Northern Hemisphere limits its view to celestial objects in the northern sky.

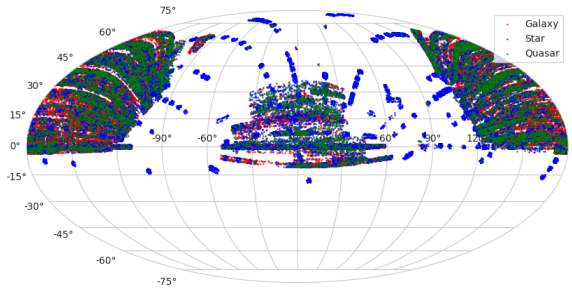

Fig. 3. Visualization of stars galaxies and quasars in the celestial coordinates.

This information is helpful because it allows astronomers to precisely locate and catalog celestial objects, enabling efficient analysis and comparison of astronomical data. Additionally, understanding the geographic bias of data collection helps researchers interpret and contextualize observations made by instruments like the SDSS.

3.5 Feature Engineering & Scaling

Feature engineering is a very vital part of this paper. Feature engineering helps create robust models that are less sensitive to noise, outliers, and changes in the dataset. It utilizes techniques such as imputation enabling us to handle missing data effectively. By replacing missing values with suitable estimates, we maintain the dataset's integrity and mitigate potential biases during model training.

There are many techniques available for feature engineering, one such technique used in this paper is Principal Component Analysis (PCA). Principal Component Analysis (PCA) is a statistical procedure that uses a technique to convert a set of correlated variables to a set of uncorrelated variables. PCA can serve two purposes in feature engineering.

- Firstly, it can be employed descriptively. By analyzing the variation explained by the components, one can compute Mutual Information (MI) scores for these components to discern which types of variation are most predictive of the target variable.
- Secondly, PCA components can be utilized directly as features. Since these components reveal the underlying structure of the data, they may offer more informative representations than the original features. One of the use cases of PCA is Dimensionality Reduction.

In this paper we will use PCA in order to reduce the dimensionality of our dataset, i.e., reduce the number of features we will be working with. This is done by preserving the features which contribute most in improving the performance of the model in classification.

In our case the dataset has 5 wavelength bands u, g, r, i, z. Using PCA on our data will decrease the amount of operations during training and testing. For efficient classification of SGQs we only need those features which help in classifying one class from the other. Using all the available features will not only hamper the performance of the model but will also add unnecessary noise and waste a lot of computational resources.

Scaling all values to be within the (0, 1) interval will reduce the distortion due to exceptionally high values and make some algorithms converge faster.

3.6 Algorithms Performance

In this paper the two algorithms used for classification are XGBoost and Random Forest Classifier. We have chosen these algorithms based on our previous paper where we compared multiple algorithms and their performance.

XGBoost, short for eXtreme Gradient Boosting, is a powerful and widely used machine learning algorithm known for its efficiency, speed, and effectiveness in various types of predictive modeling tasks, particularly in structured/tabular data. The training process in XGBoost involves iteratively adding new decision trees to the ensemble, with each tree trained to minimize a predefined objective function, typically a loss function that measures the difference between predicted and actual values. The algorithm learns by optimizing the objective function through gradient descent, where each new tree is fitted to the negative gradient of the loss function. During prediction, XGBoost aggregates the predictions from all trees in the ensemble, typically using a weighted sum, to generate the final output. The prediction for a new instance is the cumulative sum of the predictions from each individual tree.

XGBoost employs a more universal approach to tackle overfitting, leading to enhanced performance. It utilizes parallel computing, resulting in faster execution compared to traditional Gradient Boosting. Moreover, it has the capability to handle missing data and incorporates cross-validation functionality to determine the optimal number of boosting rounds in each iteration. XGBoost requires only a few parameters to be tuned in order to achieve improved outcomes.[4]. When we used XGBoost as algorithm for our model, we got the following scores:

Learning curve in machine learning is a graphical representation that shows the relationship between a model's performance on a training and validation set over a period of time, or as a function of experience (the amount of training data). This is achieved by monitoring the training and validation scores (model accuracy) with an increasing number of training samples.

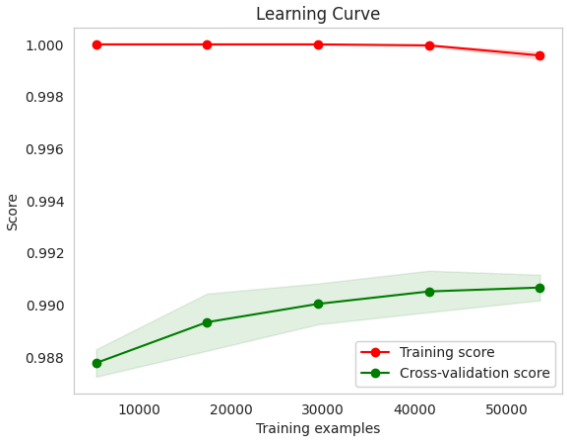

Fig. 4. Learning Curve for XGBoost algorithm

The other algorithm is Random Forest Classifier (RFC). The Random Forest classifier, a prominent ensemble learning algorithm, constructs multiple decision trees during training. It begins by generating numerous bootstrap samples from the original dataset, enabling the creation of diverse decision trees. Moreover, it employs random feature selection, which entails choosing a subset of features at random for each tree, thereby enhancing model robustness and reducing overfitting.

Each decision tree in the Random Forest is trained independently using a bootstrap sample and the selected feature subset. These trees typically grow to their maximum depth without pruning. During prediction, the ensemble combines the individual predictions of all trees using a majority voting mechanism for classification tasks or averaging for regression tasks. This aggregation of predictions contributes to the algorithm's resilience to noise and improves generalization performance. Random Forest offers various advantages, including reduced overfitting, improved generalization, and the ability to estimate feature importance. However, it may have limitations in interpretability and inference speed compared to individual decision trees.

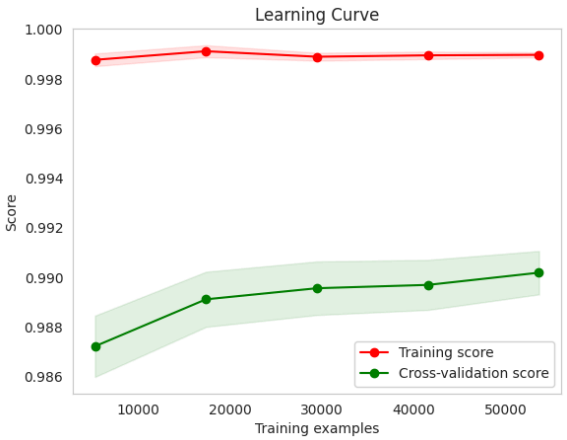

Fig. 5. Learning Curve for Random Forest Classifier algorithm

Below are the test results of the 2 algorithms i.e., XGBoost and Random Forest Classifier for classification of Stars, Galaxies and Quasars. Both the algorithms have a very similar performance and time requirement.

Models	**Prediction accuracy for train data (%)**	**Prediction accuracy for test data (%)**	**Time consumed for training (seconds)**	**Time consumed for prediction (seconds)**
XGBoost	99.91	99.03	4.526	0.35491
Random Forest Classifier	99.90	99.02	3.025	0.11499

3.7 K-Fold Cross Validation

K-fold cross-validation is a robust technique used to assess the performance of machine learning models, especially when the dataset is limited or prone to variance. The process involves dividing the dataset into k equally sized subsets, or folds. During each iteration, one of the folds is held out as the validation set while the remaining k-1 folds are used for training the model. This procedure is repeated k times, with each fold serving as the validation set exactly once. By averaging the performance metrics obtained from each iteration, k-fold cross-validation provides a more reliable estimate of the model's performance compared to a single train-test split.

One of the key advantages of k-fold cross-validation is that it maximizes the use of data for both training and validation purposes, thereby reducing the risk of overfitting. Additionally, it provides a more accurate assessment of the model's generalization ability, as it evaluates the model's performance on multiple subsets of the data. Furthermore, k-fold cross-validation helps to identify any variability in the model's performance across different subsets of the data, offering insights into its stability and robustness. Overall, k-fold cross-validation is a valuable tool for model evaluation and selection, helping practitioners make informed decisions about their machine learning algorithms.

We applied a 10-fold cross validation on XGB and RFC in order to make sure that these algorithms are not providing us with accuracies which are high just by chance. But when we applied cross validation on XGB and RFC, we found that they were indeed performing really well and their accuracies didn't change much at all.

Models	Mean of the Scores (%)	Standard Deviation
XGBoost	0.9916	0.000884
Random Forest Classifier	0.9909	0.000773

XGBoost showed a higher mean and lower standard deviation than the Scikit-Learn RFC. A high mean corresponds to a more stable performance and a low standard deviation corresponds to a smaller range of results.

3.8 Finding the best hyperparameters

Hyperparameters, also known as model hyperparameters, are external configuration settings employed by data scientists to govern the training process of machine learning models. Unlike parameters, which are internal variables automatically adjusted by the model during training, hyperparameters must be manually specified before initiating the training process. These settings play a crucial role in influencing the behavior and performance of the model, encompassing aspects such as the model's complexity, regularization strength, and optimization strategy. By carefully selecting and fine-tuning hyperparameters, data scientists can optimize model

performance and enhance its ability to generalize to unseen data.

Below table show the accuracies with optimal hyperparameters applied.

Model	Prediction accuracy WITH optimal hyperparameters (%)
Random Forest Classifier	99.01
XGBoost	99.95

4. Results

With all the above tests and iterations it is pretty clear that XGBoost and Random Forest Classifier have a very similar performance on the SDSS dataset. They came out with close by accuracies, training time and prediction time.

Models	Prediction accuracy for test data (%)	Time consumed for training (seconds)	Time consumed for prediction (seconds)
XGBoost	99.03	4.526	0.35491
Random Forest Classifier	99.02	3.025	0.11499

But XGBoost seems to perform slightly better than Random Forest Classifier. Even if the difference is in decimals. Even after applying k-fold cross validation, the XGBoost accuracy scores were slightly higher than Random Forest Classifier. This can be evidently seen below.

Models	Mean of the Scores (%)	Standard Deviation
XGBoost	0.9916	0.000884
Random Forest Classifier	0.9909	0.000773

But after hypertuning both the algorithms and comparing their accuracies shows a slight change. After hypertuning the algorithms the XGBoost in fact performed slightly worse than it performed without any hyper tuning. This may be an indication that the XGBoost without any hypertuning was classifying the SGQs at its best. Whereas Random Forest Classifier did not show any noticeable change.

Model	Prediction accuracy for test data (%)	Prediction accuracy WITH optimal hyperparameters (%)
Random Forest Classifier	99.02	99.01
XGBoost	99.03	98.95

5. Conclusions

The primary objective of this study was to evaluate the performance of two advanced machine learning algorithms, Random Forest Classifier (RFC) and XGBoost, in classifying celestial objects from the Sloan Digital Sky Survey (SDSS) Data Release 18 (DR18) dataset. Our analysis focused on determining whether fine-tuning these algorithms could yield improved classification accuracy compared to their default configurations.

The SDSS dataset, renowned for its extensive and detailed astronomical observations, provided a robust basis for this comparative study. By employing a subset of 100,000 data points, we ensured that our evaluation was both comprehensive and manageable. Our findings indicate that both RFC and XGBoost exhibit excellent performance in classifying stars, galaxies, and quasars, with prediction accuracies exceeding 99%. This underscores their efficacy and reliability in handling large, complex datasets typical in astronomical research. The application of k-fold cross-validation further validated the stability and robustness of these models, with minimal variance observed in their performance metrics.

Interestingly, hyperparameter tuning did not lead to significant improvements in the algorithms' accuracy. In fact, XGBoost's performance showed a slight decline post-tuning, suggesting that the default configuration was already well-suited for the task at hand. This highlights an important insight: for certain datasets and problems, extensive hyperparameter optimization may not always yield substantial gains and could potentially lead to overfitting or diminished performance.

Moreover, our study reinforces the value of ensemble methods like Random Forest and advanced boosting techniques like XGBoost in the realm of astronomical data analysis. Their ability to handle high-dimensional data, mitigate overfitting, and deliver accurate predictions makes them indispensable tools for researchers.

In conclusion, this research confirms that both XGBoost and Random Forest Classifier are highly effective for the multiclass classification of astronomical objects within the SDSS dataset. While hyperparameter tuning did not enhance their performance significantly, the algorithms' inherent capabilities ensured high accuracy and reliability. Future research could explore the integration of additional features, different preprocessing techniques, or hybrid models to further push the boundaries of classification accuracy. Additionally, applying these algorithms to other astronomical datasets or classification tasks could provide broader insights into their generalizability and potential applications in different contexts.

Overall, this study contributes to the growing body of literature on machine learning applications in astronomy, offering valuable perspectives on the strengths and limitations of XGBoost and Random Forest Classifier in handling large-scale astronomical data.

6. References

1. Multiclass Classification Using Random Forest Classifier. - Sermista Talla, Pavani Venigalla, Ashmitha Shaik, Meghana Vuyyuru.
2. Supervised Classification of Stars, Galaxies and Quasars using Photometric and Spectroscopic measurements from the Sloan Digital Sky Survey. - Rafid Bendimerad.
3. Credit Card Fraud Prediction Using XGBoost: An ensemble Learning Approach. - Krishna Kumar Mohbey, Mohammad Zubair Khan, Ajay Indian.
4. Johnson, H. L., Morgan, W. W. (1953) Fundamental stellar photometry for standards of spectral type on the revised system of the Yerkes spectral atlas.
5. Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines C. Elting, C. A. L. Bailer‐Jones, and K. W. Smith
6. Deep learning Approach for Classifying, Detecting and Predicting Photometric Redshifts of Quasars
7. Feature Selection Applied to Data from the Sloan Digital Sky Survey
8. Machine learning classification of stars, galaxies and quasars. Logistic, decision
9. Stellar Classification by Machine Learning. random forest support vector
10. Unsupervised star, galaxy, QSO classification Application of HDBSCAN
11. Classification and Feature Prediction of Star, Galaxies, Quasars, and Galaxy Morphologies Using Machine Learning. XGBoost, navies bayes
12. Machine Learning in Astronomy: A Case Study in Quasar-Star Classification.

13. Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars
14. Star-Galaxy Classification of Photometric Data – A Comparative Study of Machine Learning Algorithmic Models
15. Astronomical Point Source Classification through Machine Learning
16. Photometric identification of compact galaxies, stars, and quasars using multiple neural networks
17. Deep learning Approach for Classifying, Detecting and Predicting Photometric Redshifts of Quasars
18. Classification of quasars, galaxies, and stars using Multi-Modal Deep Learning
19. Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra
20. Feature Selection Applied to Data from the Sloan Digital Sky Survey
21. Machine Learning in Astronomy: A practical overview
22. Machine learning based catalogs of quasars and galaxies for cosmological studies
23. Automated physical classification in the SDSS DR10. A catalogue of candidate Quasars.