Supervised Classification of Stars, Galaxies and Quasars using Photometric and Spectroscopic measurements from the Sloan Digital Sky Survey.

RAFID BENDIMERAD[1,*]

[1] PhD. candidate, Cornell Sibley School of Mechanical and Aerospace Engineering, Upson Hall - 124 Hoy Rd, Ithaca, NY 14850.
[*] arb399@cornell.edu

Compiled May 18, 2023

This research presents a comparative study of Support Vector Machine (SVM) and Random Forest (RF) supervised learning algorithms in the classification of astronomical objects—stars, galaxies, and quasars—using photometric and spectroscopic data from the Sloan Digital Sky Survey (SDSS). A grid search in a 3-fold cross-validation scheme was employed to determine the optimal parameters for both SVM and RF models. Although the training of the SVM model was found to be approximately 50 times slower than the RF model, it exhibited a superior test accuracy of 0.940 compared to the RF's 0.823. The redshift value, measured via spectroscopy, emerged as the most significant feature for the classification process. Furthermore, a noise sensitivity test revealed that the test accuracy experienced a significant drop when the signal-to-noise ratio fell below 15.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

The Sloan Digital Sky Survey (SDSS) is a comprehensive project launched in 2000 that employs a dedicated telescope, showcased in Fig. 1, at the Apache Point Observatory located in New Mexico, United States, to deliver an unparalleled mapping of the cosmos. With a cutting-edge 120-megapixel camera and advanced spectrographs, the telescope amasses a huge amount of data, comprising multi-wavelength images and spectral details about stars, galaxies, and quasars. The survey has successfully covered more than a quarter of the sky, cataloging hundreds of millions of astronomical objects and generating spectral data for multiple millions of these objects. [1].

Throughout its operation, which has undergone multiple stages, each with its own distinct emphasis, the SDSS has been the catalyst for many ground-breaking revelations. Such discoveries encompass insights into dark energy [2], the structure of the Milky Way galaxy [3, 4], as well as the recognition of remote quasars and other scarce celestial events [5].

The classification of the celestial objects captured by the SDSS represents a fundamental task to transform the raw data into valuable information. However, due to the vast amount of data measured by the SDSS , manual classification has become unfeasible, necessitating the use of automated methods.

Machine learning techniques are widely employed with considerable success to classify data. Among these techniques, unsupervised learning using k-means clustering algorithm has been used to classify the SDSS data and showed great results [6, 7]. In the current paper, we propose to use supervised learning algorithms, namely Support Vector Machine (SVM) [8] and Random Forest (RF) [9], to classify a subset of 100,000 data points from the SDSS data base. SVM is a powerful classifier that finds an optimal hyperplane to distinguish between classes, while RF is an ensemble method that aggregates the predictions of multiple decision trees. Each algorithm has its strengths and weaknesses, making their comparative study particularly enlightening.

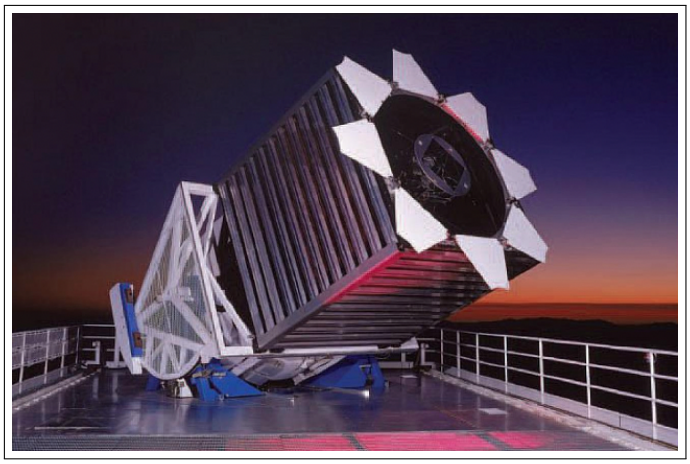

Fig. 1. SDSS telescope. Image credit: SDSS.

2. UNDERSTANDING THE DATA

A. Photometric and Spectroscopic data

The SDSS employs two primary types of measurements, namely photometric and spectroscopic, each serving distinct but complementary roles in astronomical research.

1. Photometric measurements involve capturing images of celestial objects in multiple wavelength bands. This provides a broad overview of the sky and enables the calculation of the apparent brightness and color of astronomical objects. This enables large-scale studies of the structure of the universe by cataloguing the positions and apparent magnitudes of millions of galaxies and quasars. The ugriz photometric system, utilized in the SDSS, is a sophisticated five-band system designed to obtain detailed photometric measurements across a wide spectral range. Each letter in 'ugriz' represents a specific filter: 'u' (ultraviolet), 'g' (green), 'r' (red), 'i' (near-infrared), and 'z' (further into the near-infrared). These filters are designed to cover a wavelength range from approximately 3000 to 10,000 Angstroms [10]. Fig. 2 shows a photograph of the camera of the SDSS and Fig. 3 represents the ugriz wavelength ranges.

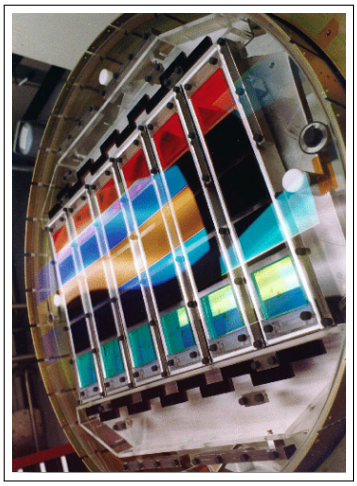

Fig. 2. Photograph of the camera of the SDSS telescope showing the ugriz filters. Image credit: SDSS.

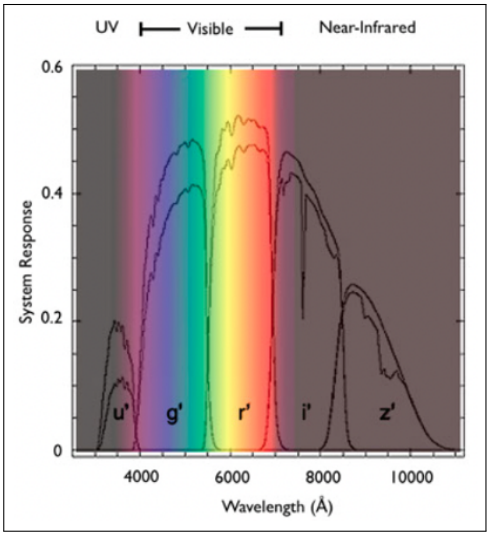

Fig. 3. Repersentatoin of the ugriz filters wavelengths. Image credit: SDSS.

2. Spectroscopic measurements, on the other hand, involve capturing the spectrum of light emitted by an object to identify its detailed chemical composition and physical properties. By splitting the light from an object into its constituent wavelengths, spectroscopy reveals a wealth of information including the velocity (through the Doppler shift), temperature, and elemental composition of the object. Spectroscopic data is particularly valuable for studying phenomena such as the expansion of the universe (via redshift measurements) and the life cycles of stars (via their elemental abundances) [11].

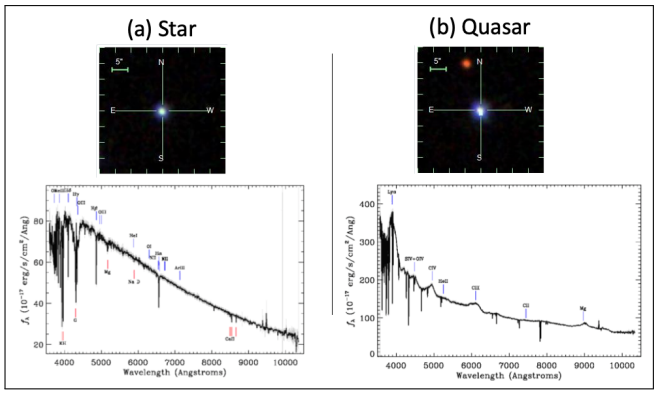

Fig. 4. Photometric images (on top) and spectroscopic measurements (on bottom) of (a) a star and (b) a quasar. Image taken from the SDSS online navigation tool.

Fig. 4 shows a comparison between the photometric and spectroscopic measurments of a star and a quasar. The use of photometry in conjunction with spectroscopy is crucial since a bright nearby star and a highly luminous distant quasar could exhibit similar brightness and color characteristics in a photometric image, making them indistinguishable based solely on these parameters. This underlines the fact that photometry, while immensely useful for characterizing the apparent properties of celestial bodies, requires the complement of other observational methods, such as spectroscopic data.

B. Representation of the data in celestial coordinates

The celestial positions of observed astronomical objects will be represented by the values of right ascension and declination. These celestial coordinates (see Fig. 5) form a fundamental reference framework for locating objects on the celestial sphere. These coordinates enable a precise and standardized assignment of locations to each observed celestial object, thereby facilitating efficient cataloging and analysis of our astronomical data.

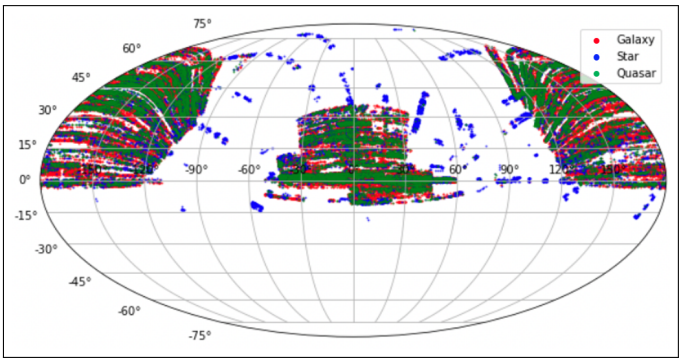

Fig. 5. Representation of the data in celestial coordinates.

The data obtained from the SDSS predominantly originates from the northern celestial hemisphere. This geographic bias is a direct consequence of the physical location of the SDSS telescope at Apache Point Observatory in New Mexico, USA. The telescope's latitude in the Northern Hemisphere restricts its access primarily to celestial objects that lie in the northern sky.

3. PRESENTATION OF THE ALGORITHMS

A. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm utilized predominantly for classification and regression analysis. The fundamental objective of SVM is to generate an optimal hyperplane that maximally segregates distinct classes of data. The hyperplane is determined by maximizing the margin, defined as the distance between the hyperplane itself and the nearest instances from each class, otherwise known as support vectors. The optimal hyperplane is obtained by minimizing the following objectinve function:

$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i\,, \tag{1}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0\,, \tag{2}$$

where $\mathbf{w}$ is the normal vector to the hyperplane, b is the bias term, ξ_i are the slack variables which allow for misclassification of difficult or noisy examples, C is the penalty parameter of the error term determining the trade-off between maximizing the margin and minimizing classification error, y_i are the class labels, and n is the total number of data points.

For non-linearly separable data, SVM employs a mathematical technique known as the kernel trick. The essential purpose of a kernel function is to transform the original feature space into a higher-dimensional space where the data becomes linearly separable, thereby allowing a hyperplane to be constructed for classification. The kernel function, denoted as $K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$, computes the inner product between two points in the feature space. Several types of kernel functions can be used, including linear, polynomial, and radial basis function (RBF), each suitable for different types of data and problem domains. The choice of kernel function and its parameters is crucial as it directly influences the performance of the SVM algorithm.

B. Random Forest

Random Forest is a supervised learning technique that constructs a multitude of decision trees during training and generates an output by aggregating the predictions of individual trees, either by mode (for classification) or mean (for regression). A Random Forest algorithm introduces an additional layer of randomness. Specifically, instead of considering all features at each candidate split during the tree-building process, Random Forest randomly selects a subset of features. This element of randomness serves to decorrelate the trees, which in turn reduces the variance of the final model without significantly increasing the bias. Moreover, this strategy enhances computational efficiency. The Random Forest algorithm's ability to adapt to various types of data, its resistance to overfitting due to randomization and aggregation, and its capacity to estimate feature importance make it a highly desirable tool for both prediction and data interpretation. However, due to its inherent complexity as a composite model, the interpretability of Random Forest can be challenging.

In the implementation of the Random Forest algorithm, several hyperparameters play crucial roles in determining the performance and the complexity of the model. The Random Forest algorithm's performance and complexity are significantly influenced by certain hyperparameters. The "Number of Estimators" parameter, which refers to the count of decision trees in the forest, is one such parameter. An increase in this number generally enhances the model's performance by reducing overfitting through the averaging of more trees, but simultaneously escalates the computational cost due to the resources required to construct and store more trees. Another important parameter is the "Maximum Depth", which determines the maximal number of levels to which each decision tree can grow. While a deeper tree can model intricate relationships by adding more nodes, it may also lead to overfitting and increased computational expense. Lastly, the "Minimum Sample Split" parameter defines the minimal number of data points a node must contain for it to be split into further child nodes. A value too low for this parameter might result in overfitting as the trees may become overly complex and capture noise in the data. Conversely, a value too high could lead to underfitting as the trees may be too simple to capture important relationships in the data. Therefore, these parameters necessitate careful tuning to strike a balance between model complexity, accuracy, and computational efficiency.

4. IMPLEMENTATION

A. Features selection

A Pearson correlation matrix is constructed for all pairs of features (see Fig. 6). Highly correlated features often carry similar information, which can be redundant for model learning. By identifying pairs of features with a correlation coefficient above a certain threshold, we can remove one feature from each pair, thereby reducing dimensionality and complexity of the model without significant loss of information. It's important to note, however, that this method only captures linear relationships and might overlook important non-linear relationships or interactions between features.

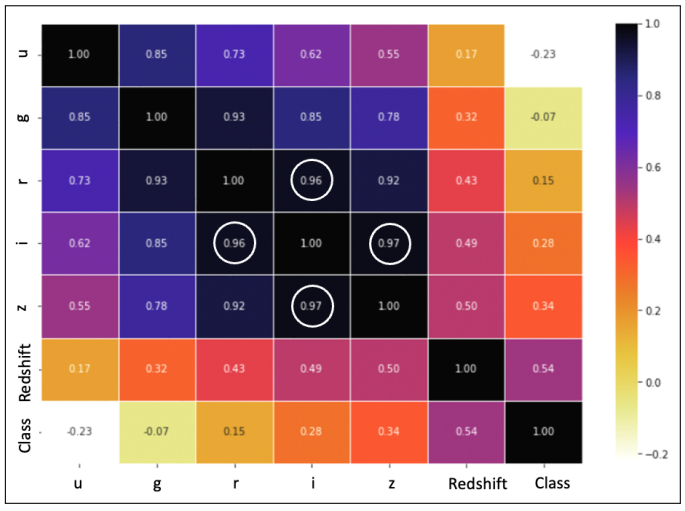

Fig. 6. Pearson Correlation Matrix of data set variables. The values above the threshold of 0.95 are circled.

For this study we choose a threshold of 0.95 and find a high correlation between the values corresponding to the red (r), near infrared (i), and infrared (z) filters. Thus, we chose to get rid of the features (i) and (z).

B. Dealing with unbalanced data

Unbalanced data refers to situations in which the classes within a classification problem are not represented equally. For instance, our data consist of 59% galaxies, 22% stars, and 19% quasars (see Fig. 7). This class imbalance can lead to biased models that favor the majority class, as the model is exposed to more examples from the majority class during training. Consequently, the model may exhibit poor predictive performance on the minority class. To address this issue, we use the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [12]. SMOTE operates by creating synthetic instances of the minority class to increase its representation, thereby mitigating the issue of the model being overly biased towards the majority class. The algorithm functions by selecting instances that are close in the feature space, drawing a line between the instances in the feature space and generating new instances along that line. Specifically, for each minority class sample, SMOTE selects 'k' nearest neighbors, picks one of them randomly, and then generates a synthetic instance at a random point between the chosen instance and its neighbor (see Fig. 8). While SMOTE can enhance the performance of machine learning algorithms on imbalanced datasets by balancing the class distribution, it may also increase the likelihood of overfitting due to the interpolation of new instances.

In our data processing workflow, we have partitioned our dataset into two distinct subsets to facilitate an effective model training and validation process. We have allocated 80% of the total data for the purpose of training, where our machine learning model will learn and tune its parameters. The remaining 20% of the data will be utilized for testing. The SMOTE algorithm will only be applied to the training set. Fig. 9 shows a bar chart of the balanced training set after applying the SMOTE algorithm.

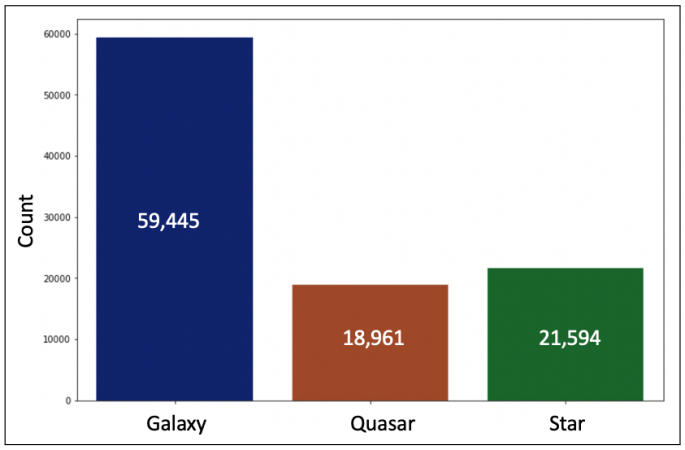

Fig. 7. Bar chart representing the count of instances for each class in the entire data set before splitting the data and applying SMOTE.

C. Grid search and 3-fold cross validation

Grid search and cross-validation are used in tandem to optimize the hyperparameters of the SVM and RF models. This combination provides a systematic and robust approach to parameter tuning, while also helping to prevent overfitting and ensuring generalization of the model to unseen data. Grid search is employed to delineate a multidimensional grid of hyperparameters, systematically training the model for each permutation of these parameters. And a 3-fold cross validation is used to appraise the performance of the model corresponding to each hyperparameter. Ultimately, the hyperparameter combination that elicits the best average performance across all folds is chosen as the optimal set of hyperparameters. The grids that represent the hyperparameters of the SVM and RF algorithms are presented in Tables 1 and 2 respectively. Whereas Fig. 10 provides a visual representation of the operational process of the 3-fold cross-validation algorithm.

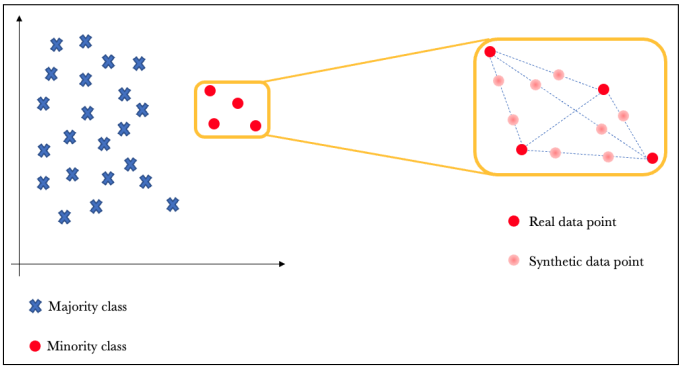

Fig. 8. Visual depiction of the SMOTE algorithm generating new synthetic data.

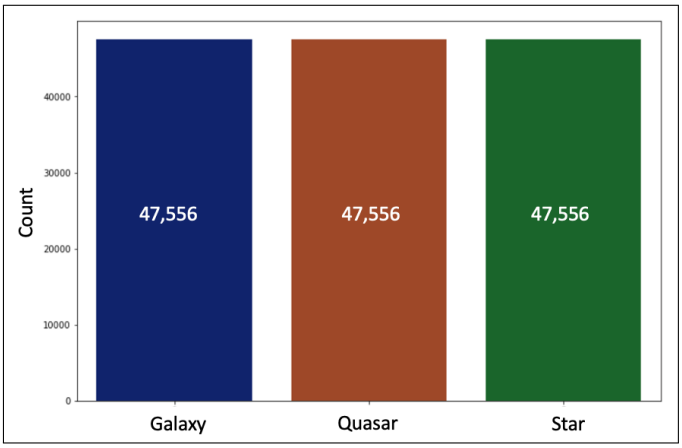

Fig. 9. Bar chart representing the count of instances for each class in the training set after applying SMOTE.

Table 1. SVM parameters used in the grid search. The best parameters are shown in bold.

Parameter	values		
C	0.1	1	**10**
Kernel	Linear	Polynomial	**rbf**
Degree	**2**	3	4

The results of the cross-validation show that the best SVM parameters are: {C = 10, degree = 2, kernel= 'rbf'}, and the best

Table 2. RF parameters used in the grid search. The best parameters are shown in bold.

Parameter	values		
Number of estimators	10	20	**50**
Maximum depth	5	10	**None**
Minimum samples split	**2**	5	10

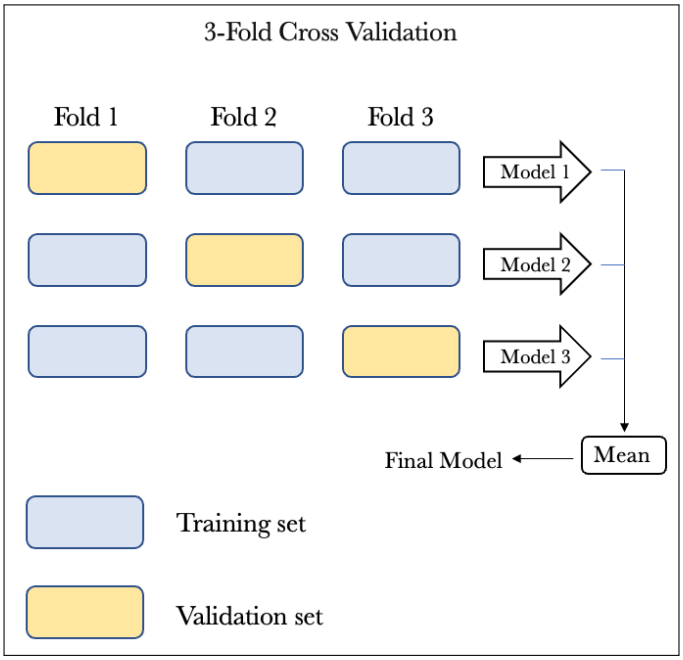

Fig. 10. Illustration of 3-Fold Cross-Validation Procedure.

RF parameters are: {$N_{estimators}$ = 50, Max depth = None, Min samples split =2}.

5. RESULTS

A. Confusion matrices

The confusion matrix is a key analytical tool used for the evaluation of the performance of classification models. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. The matrix includes values for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These four outcomes provide a more granular understanding of model performance than aggregate metrics, as they distinguish between types of correct and incorrect predictions. The confusion matrices of the SVM and RF models are shown in Fig. 11 and 12 respectively.

The confusion matrix of the SVM model shows that the True Positive (TP) rates for the predictions of galaxies, quasars, and stars are 91%, 100%, and 96% respectively. And total test accuracy of the SVM model is 0.940. On the other hand, the confusion matrix of the RF model shows that the TP rates for the predictions of galaxies, quasars, and stars are 81%, 98%, and 97% respectively. And total test accuracy of the RF model is 0.823. The SVM model is about 12% more accurate than the RF model. However, the cross validation operation took 210 minutes for the SVM and only 4 minutes for the RF model.

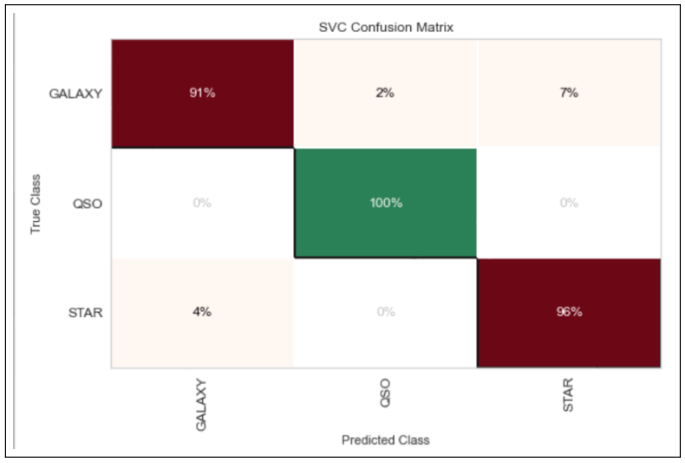

Fig. 11. Confusion Matrix of the Support Vector Machine Model.

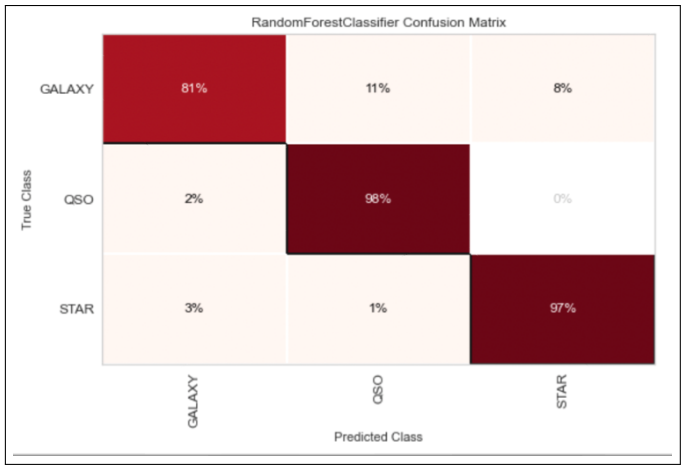

Fig. 12. Confusion Matrix of the Random Forest Model.

B. Importance plot

The importance plot, derived from the random forest algorithm, demonstrates that spectroscopic data constitute the most critical feature for the accurate classification of celestial objects (see Fig. 13). This outcome aligns with our expectations, given that stars and quasars appear remarkably similar in photometric images, rendering their differentiation challenging based solely on photometric features. However, these objects exhibit distinct spectroscopic signatures, which provide valuable insights into their intrinsic properties and physical differences. The prominence of spectroscopic data in the importance plot reaffirms its pivotal role in the successful classification of stars and quasars, underscoring the need for incorporating such data into our analytical approaches to achieve robust and accurate results.

C. Noise sensitivity test

We aim to investigate the sensitivity of the Random Forest model to the presence of noise in the data. While the initial signal inherently contains some degree of natural noise, we have further introduced synthetic random noise into both the training and test sets to examine the model's robustness under varying degrees of noise interference. The signal-to-noise ratio was systematically varied between 1 and 20 in our study (see Fig. 14).

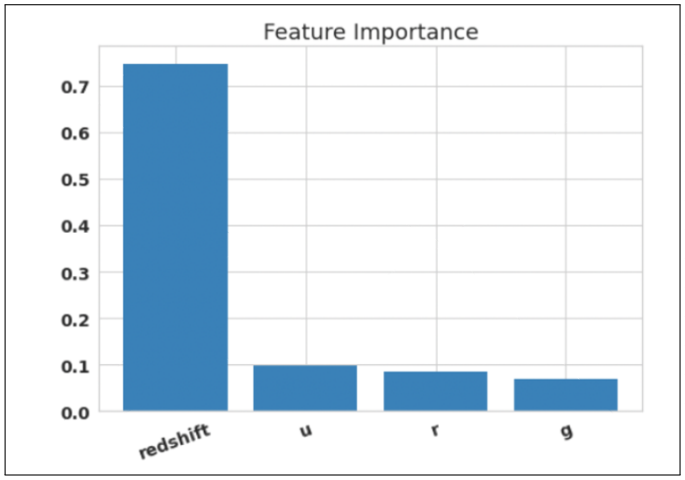

Fig. 13. Feature Importance Plot as determined by the Random Forest algorithm.

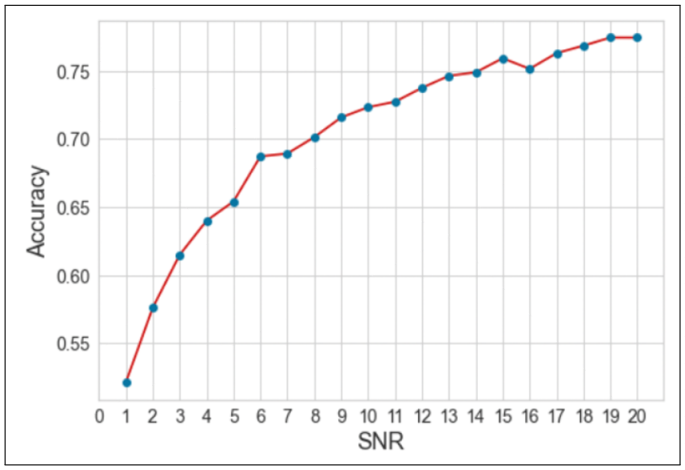

Fig. 14. Test Accuracy of the Random Forest Model versus Signal-to-Noise Ratio.

Our results indicate that the performance of the Random Forest model is significantly affected by a low signal-to-noise ratio (SNR), as demonstrated by a marked decrease in accuracy for small SNR values. As the SNR decreases, the model's capacity to discern the true signal amidst the noise diminishes, leading to an increase in misclassifications and thus a decrease in overall accuracy. However, it is noteworthy that for SNR values greater than 15, the model's accuracy converges towards 0.8. This value is approximately equivalent to the model's baseline accuracy without the introduction of additive noise, suggesting that the Random Forest model is capable of effectively filtering out noise and extracting meaningful information from the signal when the SNR is sufficiently high. This finding underscores the importance of maintaining a high SNR in our data to ensure robust and reliable performance of the Random Forest model.

6. CONCLUSION

This research has provided a comprehensive comparison of the Support Vector Machine (SVM) and Random Forest (RF) algorithms for the classification of astronomical objects using Sloan Digital Sky Survey (SDSS) data. The class imbalance issue within the data was successfully addressed using the Synthetic Minority Over-sampling Technique (SMOTE). Despite the longer training time, the SVM outperformed the RF in terms of test accuracy, thereby emphasizing the trade-off between computational efficiency and performance accuracy that is often at play in machine learning applications.

The importance of spectroscopic data is underscored, with the redshift value emerging as the most influential feature for classification. This study also illuminates the role of correlation analysis in feature selection, leading to the removal of redundant features - in this case, the near infrared (i) and infrared (z) filters due to their high inter-correlation with the red (r) filter.

Our investigation into the Random Forest model's sensitivity to noise reveals that the model's performance is significantly affected by low signal-to-noise ratios (SNR). As the SNR decreases, the model's ability to distinguish the true signal amidst the noise is compromised, leading to an increase in misclassifications and a corresponding decrease in overall accuracy. However, above an SNR of 15, the model's accuracy converges towards 0.8, roughly equal to the model's baseline accuracy in the absence of additive noise.

REFERENCES

1. https://classic.sdss.org/, "Sloan digital sky survey," (2023). Accessed: 2023-05-17.
2. Y.-C. Chen, X. Liu, W.-T. Liao, A. M. Holgado, H. Guo, R. A. Gruendl, E. Morganson, Y. Shen, K. Zhang, T. M. Abbott *et al.*, Mon. Notices Royal Astron. Soc. **499**, 2245 (2020).
3. P. Jofre and A. Weiss, Astron. & Astrophys. **533**, A59 (2011).
4. D. Carollo, T. C. Beers, M. Chiba, J. E. Norris, K. C. Freeman, Y. S. Lee, Ž. Ivezić, C. M. Rockosi, and B. Yanny, The Astrophys. J. **712**, 692 (2010).
5. D. P. Schneider, G. T. Richards, P. B. Hall, M. A. Strauss, S. F. Anderson, T. A. Boroson, N. P. Ross, Y. Shen, W. Brandt, X. Fan *et al.*, The Astron. J. **139**, 2360 (2010).
6. J. S. Almeida, J. A. L. Aguerri, C. Munoz-Tunón, and A. De Vicente, The Astrophys. J. **714**, 487 (2010).
7. J. S. Almeida and C. A. Prieto, The Astrophys. J. **763**, 50 (2013).
8. W. S. Noble, Nat. biotechnology **24**, 1565 (2006).
9. G. Biau and E. Scornet, Test **25**, 197 (2016).
10. https://www.sdss4.org/dr12/imaging/imaging_basics/, "Sdss: Understanding the imaging data," (2023). Accessed: 2023-05-17.
11. http://www.sdss3.org/dr9/spectro/, "Sdss: Spectroscopic data," (2023). Accessed: 2023-05-17.
12. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, J. artificial intelligence research **16**, 321 (2002).