STAR CLASSIFICATION

Clara Bellón **Marcin Jędrzejowski** **Santiago Suárez** **Pouria Alvarzandi**

1 Introduction

In astronomy, stellar classification is the classification of stars based on their spectral characteristics, one of the most important is the classification scheme of galaxies, quasars, and stars.

Stars are luminous celestial objects primarily composed of hydrogen and helium undergoing nuclear fusion, converting hydrogen into helium and releasing energy in the process. **Galaxies** are immense systems of stars, stellar remnants, interstellar gas, dust, and dark matter bound together by gravity. These cosmic conglomerates come in various shapes and sizes. **Quasars**, or quasi-stellar radio sources, are extremely luminous and energetic centers of distant galaxies, powered by accretion of mass onto supermassive black holes.

1.1 Description of the dataset

The data consists of 100,000 observations of space taken by the SDSS (Sloan Digital Sky Survey). Each of its rows represents a unique pattern corresponding to the observation of a celestial body. Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

- obj_ID = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS.
- alpha = Right Ascension angle (at J2000 epoch).
- delta = Declination angle (at J2000 epoch).
- u = Ultraviolet filter in the photometric system.
- g = Green filter in the photometric system.
- r = Red filter in the photometric system.
- i = Near Infrared filter in the photometric system.
- z = Infrared filter in the photometric system.
- run_ID = Run Number used to identify the specific scan.
- rereun_ID = Rerun Number to specify how the image was processed.
- cam_col = Camera column to identify the scanline within the run.
- field_ID = Field number to identify each field.
- spec_obj_ID = Unique ID used for optical spectroscopic objects.
- class = object class (galaxy, star or quasar object).
- redshift = redshift value based on the increase in wavelength.
- plate = plate ID, identifies each plate in SDSS.
- MJD = Modified Julian Date, used to indicate when a given piece of SDSS data was taken.
- fiber_ID = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation.

From these parameters we will use **u**, **g**, **r**, **i**, **z** and **redshift** for the input, as they constitute the spectral characteristics from which the classification can be derived. The other parameters are not relevant for the classification, for example, alpha and delta represent the position of the celestial body in the celestial dome. On the other hand, the desired outputs will be given by the **class** column.

It must be highlighted that this dataset is unbalanced, which means that there is a different number of samples for each class. This can affect the precision of the model as it will be shown later.

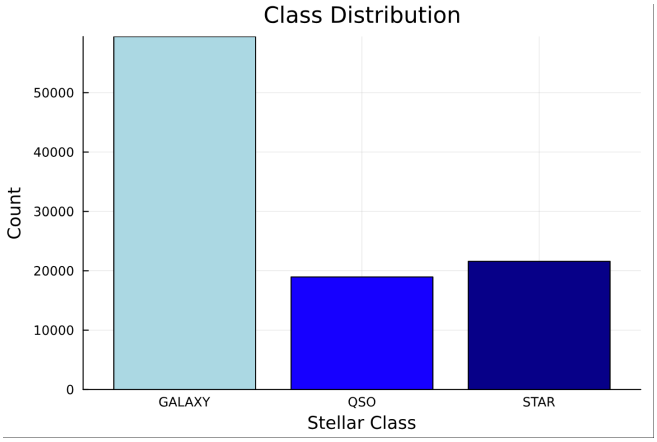

Figure 1: Distribution of the observations among the classes

1.2 Justification of the metrics

1.3 Explanation of the code structure

1.4 Bibliographic analysis

The article presents a study on classifying stars, galaxies, and quasars (QSOs) using data from the LAMOST Data Release 5 and machine learning techniques. It explores the effectiveness of four algorithms: k-nearest neighbor (KNN), decision tree (DT), random forest (RF), and support vector machine (SVM) in classifying these astronomical objects. The study finds that RF and SVM are more accurate than KNN and DT in star classification. The research highlights the challenges and potential of using machine learning for astronomical classification, especially in handling data imbalance and feature similarity among different classes.

The article titled "Automated physical classification in the SDSS DR10: A catalogue of candidate Quasars" investigates the use of machine learning, specifically the Multi Layer Perceptron with Quasi Newton Algorithm (MLPQNA), for classifying

astronomical objects based on Sloan Digital Sky Survey (SDSS) Data Release 10. The study focuses on classifying objects into galaxies, quasars, and stars using photometric data alone. The research demonstrates that while distinguishing AGN from normal galaxies using only photometric data is challenging, MLPQNA effectively separates the three main classes. The MLPQNA method achieved an overall efficiency of 91.31% and a QSO class purity of about 95%. The resulting catalogue includes approximately 3.6 million candidate quasars/AGNs, with half a million flagged as robust candidates. The study underscores the potential of machine learning in automating the classification of large astronomical datasets.

The article "Identifying Galaxies, Quasars, and Stars with Machine Learning: A New Catalogue of Classifications for 111 Million SDSS Sources without Spectra" presents a study where machine learning, specifically an optimized random forest classifier, was applied to astronomical data. The researchers trained the model using 3.1 million spectroscopically labeled sources from the Sloan Digital Sky Survey (SDSS) and applied it to classify 111 million previously unclassified sources. The new catalogue includes 50.4 million galaxies, 2.1 million quasars, and 58.8 million stars. The study also explored the effects of class imbalance on the model and the potential of transfer learning for fainter sources. The research demonstrates the power of machine learning in astronomical classification and contributes significantly to the field by categorizing a vast number of celestial bodies.

2 Development

2.1 First approach

In this initial approach to the star classifier problem, we have aimed to assess the impact of unbalanced data on the model performance, so we leave the distribution of the dataset as it is. For this experiment, we have utilized 10% of the samples from the dataset, employing normalization techniques and crossvalidation. The weak models chosen for evaluation include K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). These models were then combined into an ensemble model using Random Forest to enhance the overall predictive power.

We will try different methods of normalization and several numbers of folds in crossvalidation to see its impact on the metrics of the final model.

2.1.1 Min-max normalization, 5 folds.

First, for each weak model, we find the most suitable hyperparameters based on the results of the calculated metrics.

For example, the value of the metrics for the decision trees with maximum depths 3, 5, 10, 20, 50, 100 are:

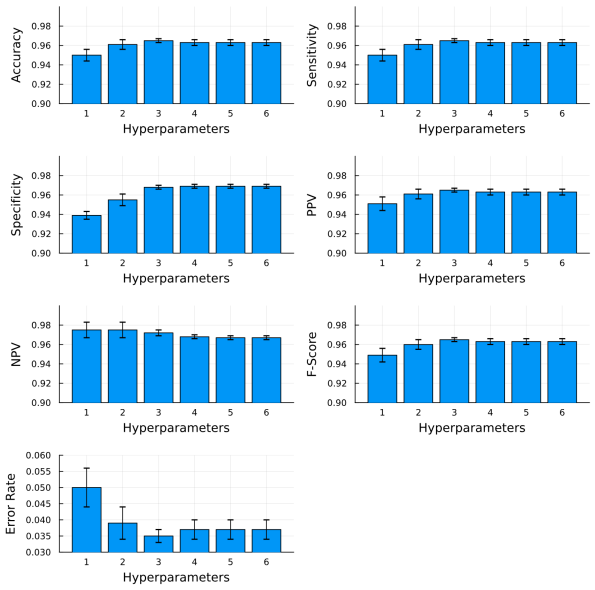

Figure 2: Metrics on decision trees with different hyperparameters.

We select the most suitable hyperparameters based on the value of the accuracy, in this one, following that rule we select the third combination of hyperparameters.

We do this same process to find the best hyperparameters for the kNN, SVM and ANN models.

Then, we train each one of the weak models with the selected hyperparameters in the whole dataset and calculate their metrics based on their predictions for the test dataset.

Finally, we ensemble the final Random Forest model using the weakmodels we just trained and calculate the metrics for this final model:

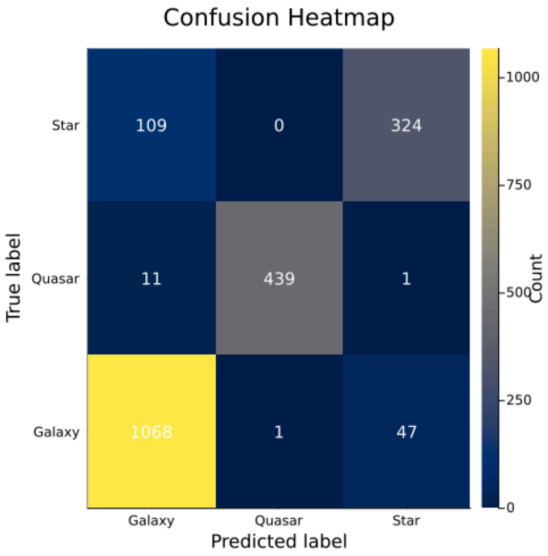

Figure 3: Heatmap for the confusion matrix of random forest.

Although, we can observe that in this heatmap the only yellow square is the one corresponding to predicting galaxies correctly, we must have in mind that the dataset is unbalanced and there is a much higher percentage of galaxy observations.

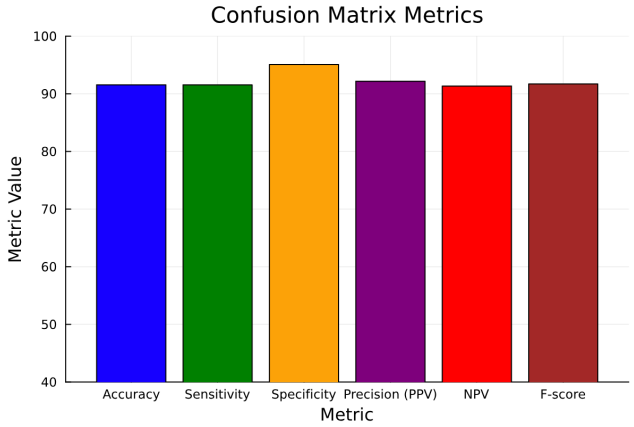

Figure 4: Random forest's metrics.

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	94.15	94.15	94.32	94.08	95.66	94.08
kNN	64.65	64.65	75.58	71.75	70.18	66.30
SVM	65.15	65.15	90.99	82.68	72.40	64.87
ANN	91.0	91.0	96.76	92.56	89.75	91.25
Ensemble	91.55	91.55	95.07	92.18	91.35	91.72

The Decision Tree model outperforms the others in almost every metric except for specificity

2.1.2 Min-max normalization, 10 folds

We use the exact same process as before but using 10 folds instead of 5 for crossvalidation.

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	94.15	94.15	94.32	94.08	95.66	94.08
kNN	64.65	64.65	75.58	71.75	70.18	66.30
SVM	80.9	80.90	93.37	86.54	81.29	81.36
ANN	91.0	91.0	96.76	92.56	89.75	91.25
Ensemble	92.05	92.05	92.78	91.92	93.64	91.97

2.1.2. Zero mean normalization, 5 folds

This time, in the preprocessing of the model we apply zero mean normalization to our data.

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	74.75	74.75	64.79	81.29	90.87	65.72
kNN	94.1	94.1	94.43	94.10	95.33	94.09
SVM	94.75	94.75	93.95	94.82	96.96	94.72
ANN	83.90	83.90	78.41	86.36	93.08	81.85
Ensemble	75.75	75.75	65.57	82.16	92.05	66.67

2.1.2. Zero mean normalization, 10 folds

We use the exact same process as before but using 10 folds instead of 5 for crossvalidation.

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	74.75	74.75	64.79	81.29	90.87	65.72
kNN	94.1	94.1	94.43	94.10	95.33	94.09
SVM	94.75	94.75	93.95	94.82	96.96	94.72
ANN	80.7	80.70	73.61	84.25	92.19	76.97
Ensemble	75.6	75.6	65.47	82.02	91.71	66.53

2.2 Second approach

In this initial approach to the star classifier problem, we have aimed to assess how well is Majority Voting ensemble method. As seen in the analysis of the data, the dataset is unbalanced; in this approach, the data is balanced using the undersampling method, which consists of selecting randomly, from all the avaliable data, the same ammount of patterns classified as stars, galaxies and quasars. For this experiment, we have utilized 10% of the samples from the balanced dataset.

The weak models chosen for evaluation include K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). These models were then combined into an ensemble model using Majority Voting to enhance the overall predictive power.

2.2.1. Min-max normalization, 5 folds

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	92.71	92.71	96.46	92.81	96.44	92.67

kNN	70.47	70.47	84.85	71.46	86.77	67.99
SVM	90.51	90.51	95.16	91.11	95.59	90.42
ANN	70.21	70.21	84.83	83.28	87.62	70.47
Ensemble	92.71	92.71	96.32	92.69	96.48	92.64

2.2.2. Zero mean normalization, 5 folds

This time, in the preprocessing of the model we apply zero mean normalization to our data.

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F Score**
Dec. tree	95.96	95.96	97.99	95.94	98.03	95.94
kNN	93.76	93.76	96.83	93.77	96.89	93.74
SVM	95.96	95.96	97.98	95.94	98.03	95.93
ANN	95.69	95.69	97.88	95.70	97.85	95.69
Ensemble	96.57	96.57	98.30	96.56	98.32	96.56

2.2.3. Zero mean normalization, 10 folds

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F Score**
Dec. tree	95.96	95.96	97.99	95.94	98.03	95.94
kNN	93.76	93.76	96.83	93.77	96.89	93.74
SVM	95.96	95.96	97.98	95.94	98.03	95.93
ANN	95.69	95.69	97.88	95.70	97.85	95.69
Ensemble	96.57	96.57	98.30	96.56	98.32	96.56

2.3 Third approach

In this approach, we distinguish our star classification strategy by incorporating Principal Component Analysis (PCA). By applying PCA, we aim to enhance model performance by reducing the dimensionality of our input features, focusing on retaining 95% of the data's variability. This significant change is expected to bring new insights compared to our previous methods.

In this iteration, the ensemble model has undergone a notable shift from our previous strategies, transitioning towards a soft voting ensemble technique. This approach considers the probability estimates from each of our carefully selected models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). By averaging these probabilities, the soft voting ensemble provides a more nuanced aggregation of predictions, which could lead to more accurate and reliable classification results, especially given the balanced and dimensionally optimized nature of our dataset.

Following the same process as before, we get the best parameters among the options we try and define a voting classifier model.

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F Score**
Dec. tree	73.37	73.37	86.62	73.27	86.59	73.28
kNN	87.79	87.79	93.79	88.03	94.09	87.63
SVM	92.09	92.09	96.00	92.25	96.19	91.96
ANN	63.97	63.97	82.01	74.43	83.60	60.02
Ensemble	82.43	82.43	91.17	82.69	91.15	82.35

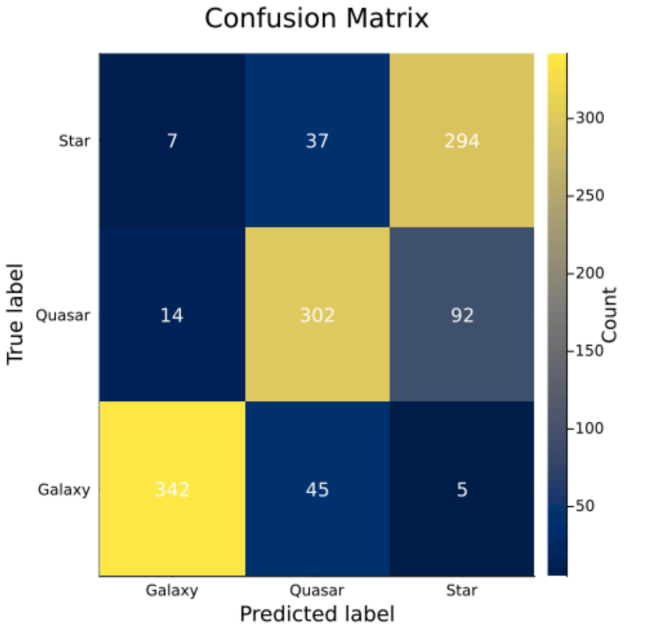

Figure 5: Heatmap for the confusion matrix of the voting classifier.

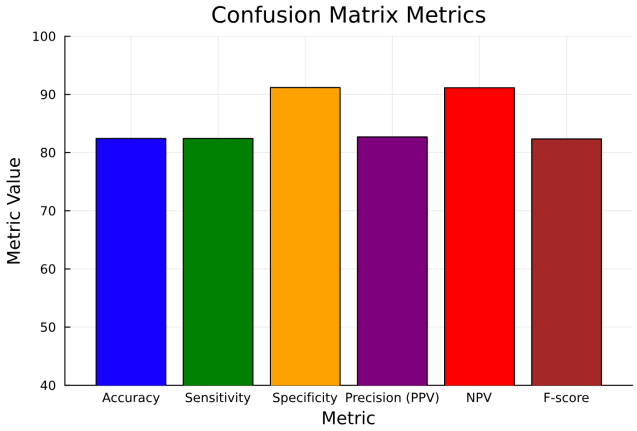

Figure 6: Voting classifier's metrics.

2.4 Fourth approach

This approach to star classification introduces a key modification: the use of unbalanced data, contrasting the previous method which employed a balanced dataset. Additionally, we have reduced the dataset size to 5% of the total data. This reduction addresses the increased execution time that unbalanced data typically demands, as dealing with disproportionate class distributions can complicate and prolong the training process. Despite the reduced dataset size, we continue to apply Principal Component Analysis (PCA), retaining 95% of the data's variability to maintain the essence of the dataset while reducing dimensionality.

In this iteration, our ensemble model adopts a weighted majority voting technique. This method differs from our previous soft voting strategy by assigning weights to each model's predictions, reflecting their respective performance levels. Our ensemble comprises the same set of models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT). However, each model's influence in the final prediction is now adjusted based on its demonstrated accuracy and reliability. This weighted approach is designed to leverage the strengths of each model more effectively, potentially leading to improved classification accuracy in an unbalanced data scenario.

	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score
Dec. tree	87.5	87.5	92.57	88.49	87.11	87.66
kNN	92.0	92.0	94.84	92.54	91.54	92.10
SVM	93.10	93.10	96.53	93.83	91.95	93.19
ANN	94.0	94.0	96.89	94.56	93.08	94.08
Ensemble	94.20	94.20	96.94	94.71	93.33	94.27

2.5 Best model

Based on the hyperparameters provided for each model, the best configurations according to accuracy are:

Artificial Neural Network (ANN): The best-performing ANN model uses the architecture [50, 30] with 'tanh' activation, a learning rate of 0.01, a validation fraction of 0.1, and a maximum of 10000 iterations.

Support Vector Machine (SVM): The SVM model that performed best had the 'linear' kernel with a degree of 7, a C value of 10.0, and 'scale' gamma setting.

Decision Tree: The best decision tree model had a maximum depth of 10.

K-Nearest Neighbors (KNN): The KNN model that yielded the best results had 5 neighbors, indicating a moderate level of complexity.

Ensemble Model: The stacking ensemble now combines Decision Tree, kNN, ANN, and SVM models with a final estimator using a Random Forest classifier.

The Decision Tree model now outperforms the others with the highest accuracy of 94.15%, slightly outperforming the Ensemble model. It appears to offer a well-balanced compromise between bias and variance, showcasing superior performance in this evaluation and the power of decision tree methods. While accuracy remains a key metric, it's crucial to consider other metrics like sensitivity or specificity, where the best results were obtained by the ANN and Decision Tree models.

3 Final discussion

In the first approach the values of the metrics are lower than for other approaches because of the unbalanced dataset; in particular, they are lower than those of the second approach, which only difference with the first is the balanced dataset. This shows the importance of a well balanced dataset in the training of our models.

As mentioned before, the first 2 approaches use only the 10% of the dataset, this is purely a matter of time as the checking of the parameters and training of the models takes around 80 minutes for the first approach and 65 for the second. Also we should note that each of this approaches repeats the process several times.

REFERENCES

[1] Wen Xiao-Qing, Yang Jin-Meng. 2021. *Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches* (303-311). Science of School, Nanchang University, Nanchang, 330031, PR China.

[2] M. Brescia, S. Cavuoti, G. Longo. 2015. *Automated physical classification in the SDSS DR10. A catalogue of candidate Quasars* (1-13). Astronomical observatory of Capodimonte, Napoli, Italy. Department of Physics, University Federico II, Napoli, Italy.

[4] A. O. Clarke, A. M. M. Scaife, R. Greenhalgh and V. Griguta. 2020. *Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra.* Jodrell Bank Centre For Astrophysics.