Shining Bright: AI-Driven Classification of Stars, Galaxies, and Quasars

Ananya Bandyopadhyay[1*], *Rupam Jash*[2], *Sourav Chowdhury*[3] *and Suparna Roychowdhury*[3]

[1]*National Institute of Technology Rourkela*
[2] *Observatoire de Paris - PSL*
[3]*St. Xavier's College (Autonomous) Kolkata*

**Email: bandyopadhyay908@gmail.com*

Abstract

Stars, galaxies, and quasars are fascinating celestial objects that play pivotal roles in our understanding of the universe. Stars are luminous spheres of plasma, whereas galaxies are vast systems which contain stars, quasars, gas, dust, and dark matter held together by gravity and quasars, short for "quasi-stellar radio sources", are extremely bright and distant celestial objects and can outshine entire galaxies. Large galaxy surveys, like Solan Digital Sky Survey(SDSS), aim to observe and catalogue a vast number of celestial objects within a given region of the sky. Thus, the first step in the study of these objects is to classify them. This classification provides a systematic framework for organizing and studying these objects. It allows astronomers to make meaningful comparisons, identify relationships and connections, and ultimately gain insights into the fundamental processes that shape our universe. The conventional classification process for these objects is known to be laborious and time-consuming. However, by leveraging AI and machine learning techniques, the task becomes much faster. The objective of this paper is to classify these objects based on data from SDSS by employing supervised machine learning algorithms and developing neural networks to compare their performance in terms of accuracy and time required for classification. The dense neural network, we built, showed the best results in terms of both accuracy and time. We saw that decision-tree-based ensemble methods performed much better than the other algorithms, the details of which are discussed in the manuscript.

Keywords: *Astronomical data survey***,** *Classification, Artificial Neural Network, Supervised machine learning algorithms, Confusion matrix.*

1. Introduction

For millennia, humans have been fascinated by the cosmos and have sought to understand the mysteries of the universe. This curiosity has led to the development of astronomy, cosmology, and various scientific and philosophical disciplines. The desire to explore and discover has driven humanity to venture beyond its immediate surroundings, leading to remarkable achievements such as the exploration of space and the development of advanced telescopes and instruments that allow us to peer deeper into the universe. Observing the universe through telescopes has revealed a vast array of celestial bodies, many of which might appear similar to the naked eye but are actually diverse and distinct objects with unique properties. Therefore, identifying and classifying these celestial bodies correctly is a crucial task in astronomy.

The utilization of Artificial Intelligence (AI) for the classification of stars, galaxies, and quasars offers several compelling advantages that complement existing classification methods. Modern astronomical surveys generate vast amounts of data, containing information about millions of celestial objects. AI can process and analyze these extensive datasets quickly and efficiently, enabling the classification of a colossal number of objects in a shorter timeframe. It can effectively handle complex data and identify subtle patterns that might be challenging for human observers or traditional methods to discern.

Here we look into data from one such survey namely the Solan Digital Sky Survey (SDSS)[7].The Sloan Digital Sky Survey (SDSS) is a comprehensive astronomical survey that has provided a wealth of data on stars, galaxies and quasars. We use this data to train machine learning models to accurately classify and identify these objects based on their observed properties.

2. The shining bodies

In this section, we delve into a brief overview of stars, galaxies, and quasars to gain insights into these fascinating celestial objects. These objects are not only rich laboratories where we can test our physical theories but are also extremely relevant in understanding the diverse nature of objects formed in the universe and their hierarchy.

2.1 Stars

Stars are luminous celestial objects composed primarily of hydrogen and helium that undergo nuclear fusion in their cores, releasing energy in the form of light and heat. A star's luminosity is a measure of the total energy it radiates into space, encompassing all wavelengths of light. Their luminosities range from 0.001 solar luminosity to thousands or even millions of times the solar luminosity. Their immense gravitational pressure and temperature sustain nuclear reactions, converting hydrogen into helium and releasing photons. This process creates a delicate balance between the outward pressure from the fusion reactions and the inward gravitational pull. Stars come in various sizes, from smaller dwarf stars like red dwarfs to massive giants and supergiants. A star's size correlates with its stage of evolution; larger stars typically evolve faster and exhaust their nuclear fuel more quickly. This diversity in size and luminosity contributes to the stunning variety of stars observable in the night sky, each offering valuable insights into the life cycles and dynamics of the universe.

2.2 Quasars

Quasars, short for Quasi-Stellar Objects, resemble faint stars in telescopic images, but their true essence lies in their nature as the energetic cores of distant galaxies, powered by supermassive black holes. These black holes, with masses millions to billions of times that of our Sun, actively accrete matter from their surroundings, creating a swirling accretion disk of superheated gas. This process releases an incredible amount of energy across the electromagnetic spectrum, from radio waves to X-rays, resulting in the immense luminosity that characterizes quasars. In fact, quasars can outshine entire galaxies, making them visible across billions of light-years. These distant celestial objects reach luminosities of over 100 trillion times the solar luminosity. Quasars play a particularly intriguing role as cosmic beacons, allowing astronomers to peer back in time to the early universe. Because light takes time to travel across vast cosmic distances, observing a quasar a billion light-years away provides a snapshot of the universe as it existed a billion years ago. Quasars offer a window into the conditions of the cosmos shortly after the Big Bang, helping scientists understand the evolution of galaxies, the intergalactic medium, and the nature of space-time itself.

2.3 Galaxies

Galaxies, the majestic enclaves of the cosmos, stand as cosmic conglomerates that house countless stars, interstellar matter, and mysteries of the universe. They come in a stunning array of forms, from the graceful spirals that wind like cosmic whirlpools to the smooth ellipses and the irregular splatters of stars and gas. Galaxies vary in size and luminosity, from the dwarf galaxies, modest in their stellar population, to the luminous giants that shine brilliantly across the cosmic canvas. Their typical luminosity are about a trillion times the solar luminosity. Each galaxy carries a unique tale of cosmic history, shaped by gravitational interactions, star formation, and the presence of supermassive black holes at their cores. Studying galaxies not only unveils their secrets but also provides insights into the grand cosmic narrative, from the birth of stars to the weaving of the cosmic web.

3. Dataset description

The Sloan Digital Sky Survey (SDSS) is a prominent multi-spectral imaging and spectroscopic redshift survey, carried out using a dedicated 2.5-meter wide-angle optical telescope situated at the Apache Point Observatory in New Mexico, United States. The survey has proven to be a cornerstone in modern astronomy, facilitating the exploration of the cosmos on a grand scale. Here we have used the Data Release 17 (DR17)[8]. Released in 2020, DR17 builds upon the legacy of previous data releases by providing an even more comprehensive and detailed dataset. Among its spectroscopic attributes, SDSS employs key keywords for precise data characterization, for example their coordinates (RA, DEC), magnitudes and redshifts:

- RA Keyword: This field denotes the right ascension of the target observed, representing its celestial coordinate on the sky in a non-negative real floating-point format.
- DEC Keyword: The declination of the target is captured in this field, provided as a real floating-point number to accurately indicate its location in the celestial sphere.
- MAGTYPE Keywords: These keywords correspond to the magnitudes of the target in different spectral bands. The values, real floating-point numbers ranging from 0 to 100, represent the intensity of light received through filters, specifically u, g, r, i, and z.
- Redshift Keyword: Represented as a real floating-point number, this field encapsulates the redshift of a target. The redshift is a pivotal astronomical metric indicating the object's distance from us, derived from the analysis of its spectral lines.

SDSS data encompasses a multitude of other attributes, albeit of lesser significance for the primary task of classification based on spectroscopic data. These include “Objid” Keyword, “Run”, “Rerun”, “Camcol”, “Field”, “Specobjid”, “Plate”, “Mjd”, “Fiberid”. These carry unique identifiers or technical specifications.

4. Data preprocessing

In the pursuit of exploring machine learning models for data analysis, our focus has been primarily directed towards a subset of parameters, namely "u", "g", "r", "i", "z", "class", "redshift", and "plate". This choice was made after careful consideration and deliberate omission of some parameters, including "objid", "ra", "dec", "run", "rerun", "camcol", "field", "fiberid", and "specobjid".

These variables were not arbitrarily excluded from our study; rather, it was driven by the nature of our research objectives and the specific context in which we were working. "objid", "ra", and "dec" are concerned with identifiers and spatial coordinates, respectively, which may not be pertinent to our current investigation. Similarly, "run", "rerun", "camcol", and "field" are related to the observational setup and do not directly contribute to the features we are investigating. "fiberid" and "specobjid" are largely concerned with technical aspects of data collecting and do not hold immediate significance to our exploration. "fiberid" and "specobjid" are largely concerned with technical data collecting aspects and are not immediately relevant to our investigation. Modified Julian Date (mjd) might be important for temporal analysis, but the study's current focus could make its omission acceptable.

We queried a significant sample size of three hundred thousand data samples from SDSS. The distribution of this data is shown below (Fig:1). Of this data 80% went for training and 20% for the subsequent model assessment.

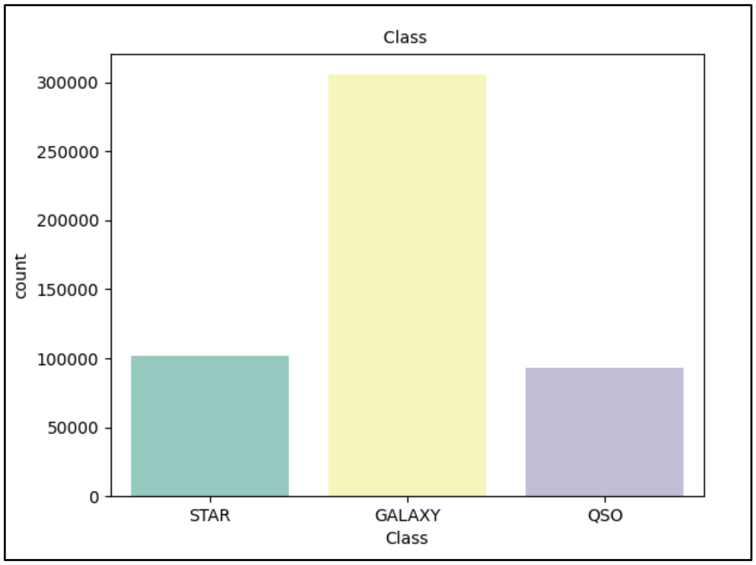


FIG 1. Distribution of objects in queried SDSS data

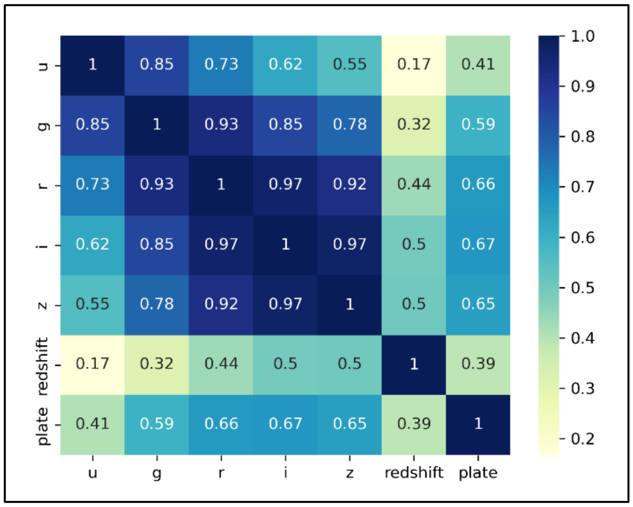
In this particular case, a Pearson correlation matrix has been used to examine the relationships between the predictor variables "u", "g", "r", "i", "z", "redshift", and "plate". The Pearson correlation coefficient measures the linear correlation between two variables, indicating the strength and direction of their relationship.

In general, we know that if the Pearson correlation coefficients r greater than 0.8, then they are highly correlated. From Fig: 2 we have found predictor variables in our analysis utilizing the Pearson correlation matrix that have Pearson correlation coefficients (r) greater than 0.8. These high correlation values signify strong linear relationships between the variables.

Below are the pairs of predictor variables with Pearson correlation coefficients greater than 0.8:

1. Variable "g" and Variable "u" (r = 0.85)
2. Variable "r" and Variable "g" (r = 0.93)
3. Variable "i" and Variable "g" (r = 0.85)
4. Variable "i" and Variable "r" (r = 0.97)
5. Variable "z" and Variable "r" (r = 0.92)
6. Variable "z" and Variable "i" (r = 0.97)

From the correlation matrix provided below (Fig: 2), it is evident that several variables exhibit strong correlations. This observation suggests the presence of multicollinearity within the system.



	u	g	r	i	z	redshift	plate
u	1	0.85	0.73	0.62	0.55	0.17	0.41
g	0.85	1	0.93	0.85	0.78	0.32	0.59
r	0.73	0.93	1	0.97	0.92	0.44	0.66
i	0.62	0.85	0.97	1	0.97	0.5	0.67
z	0.55	0.78	0.92	0.97	1	0.5	0.65
redshift	0.17	0.32	0.44	0.5	0.5	1	0.39
plate	0.41	0.59	0.66	0.67	0.65	0.39	1

FIG 2. Pearson correlation heat-map of the SDSS data.

In the pairplot (Fig: 3), the diagonal components consist of univariate density plots with respect to the class. Analyzing these diagonal plots reveal that the individual graphs are dissimilar. The lower triangular matrix in the plot displays scatter plots depicting the relationship between pairs of predictor variables in relation to the class. It becomes apparent that when one of the predictor variables is held constant, the redshift for Quasi-Stellar Objects (QSOs) is consistently higher than that for galaxies and stars. Thus, redshift is an important machine learning variable for our algorithm.

However, this data is imbalanced. Hence, we used Synthetic Minority Oversampling Technique (SMOTE) which is a statistical method employed to equilibrate the representation of different classes within the test dataset. This technique operates by creating fresh instances derived from the existing minority class cases provided as input. The distribution of this data is shown below (Fig:4).

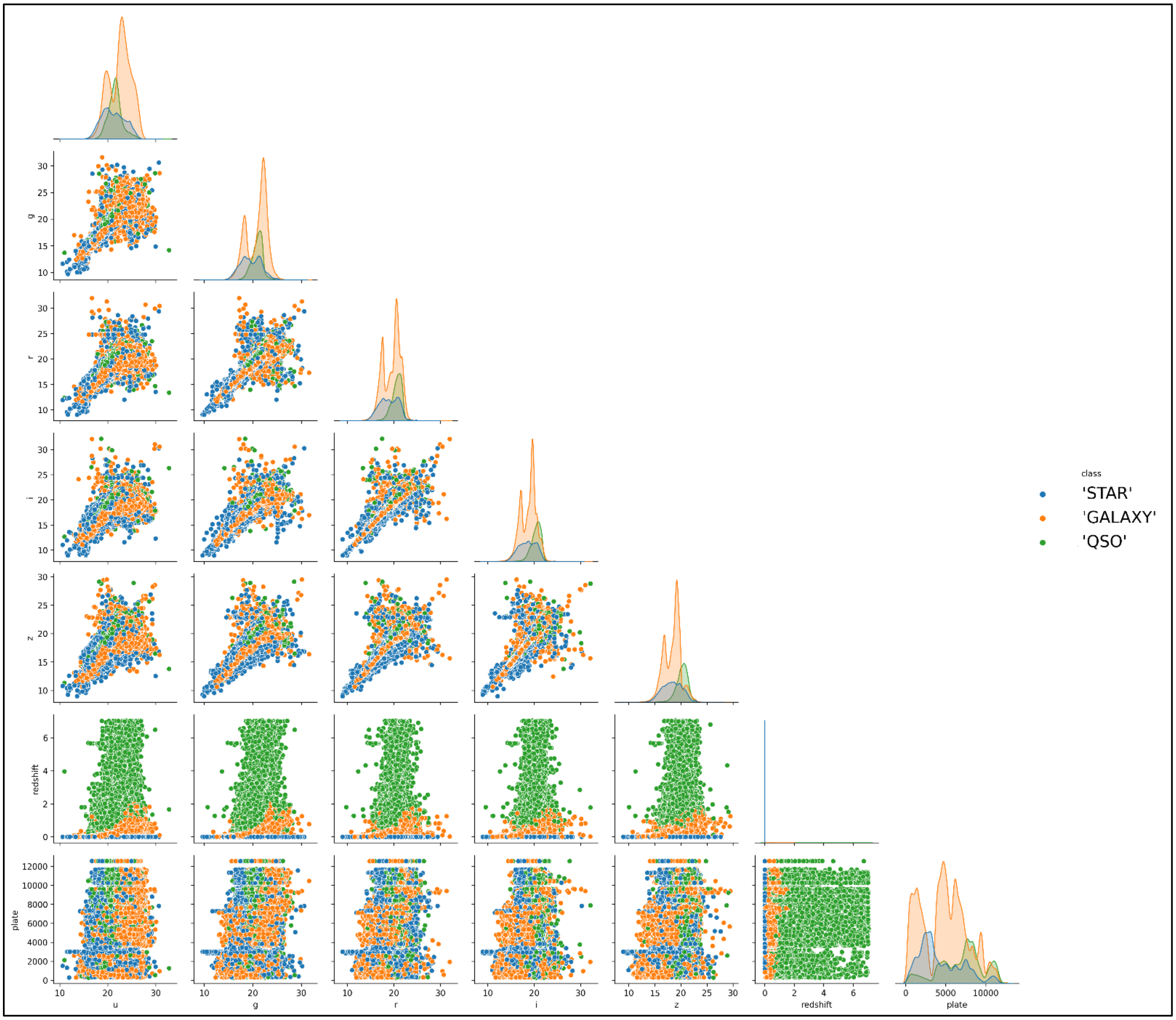


FIG 3. Piecewise pair-plot between predictor variables of the SDSS data

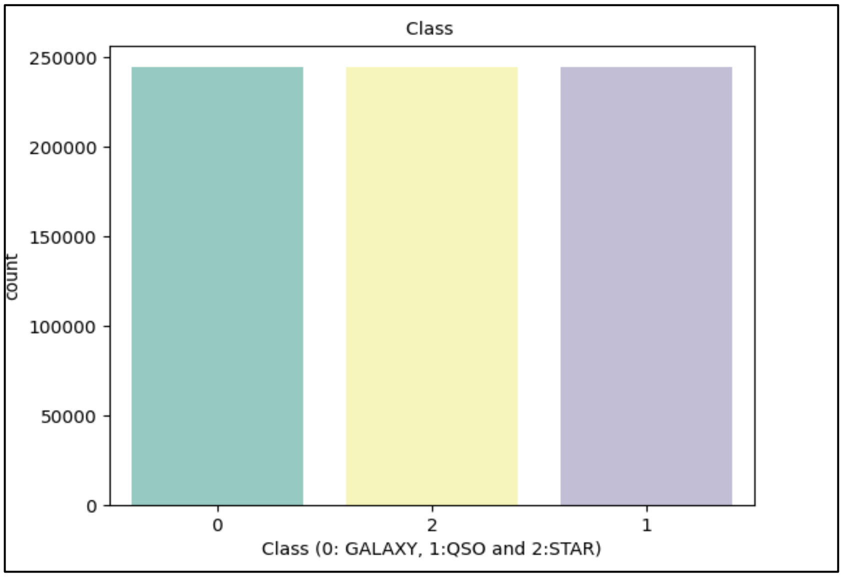


FIG 4. Distribution of test dataset after using SMOTE

5. Models Explored

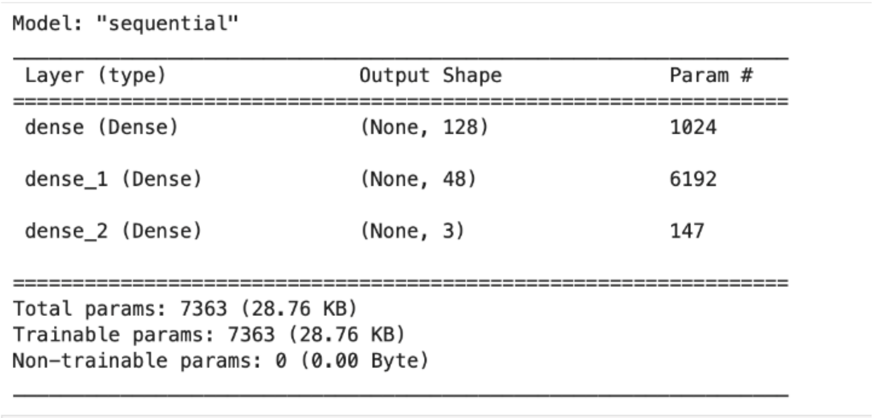
In this section, we explore the algorithms that we have used for this study.

5.1 Artificial Neural Network

An Artificial Neural Network (ANN) is a type of machine learning model inspired by the structure and functioning of the human brain. An ANN consists of interconnected nodes, called neurons, organized into layers. Each neuron takes inputs, performs a weighted sum of these inputs, adds a bias term, and then passes the result through an activation function. Layers are sets of neurons that are organized in a specific order. The input layer receives the data, hidden layers process the data, and the output layer produces the final result.

Here we have built a sequential neural network with 3 dense layers whose architecture comprises 128 nodes in the first layer, 48 nodes in the second layer, and finally, 3 nodes in the third layer.

It makes obvious and practical sense to use Rectified Linear Unit (ReLU) activation functions for the top two levels. ReLU activation allows the network to include non-linearity while maintaining computational efficiency. The third layer, comprising 3 nodes, adopts the sigmoid activation function since the network is applied to a problem involving three-class classification. The model summary is shown below (Fig: 3).



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               1024

 dense_1 (Dense)             (None, 48)                6192

 dense_2 (Dense)             (None, 3)                 147

=================================================================
Total params: 7363 (28.76 KB)
Trainable params: 7363 (28.76 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

FIG 5. Model Summary

5.2 Supervised Machine Learning Algorithms

1. Naïve Bayes Classifier
   The Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' theorem and the assumption of conditional independence among features.

   We have employed Gaussian Naive Bayes (Gaussian NB()) from scikit-learn here, which that assumes that features follow a Gaussian (normal) distribution.

2. Support Vector Machine Classifier
   Support Vector Machine (SVM) works by finding a hyperplane that best separates different classes of data points. The goal is to maximize the margin, i.e., the distance between the hyperplane and the nearest data points from each class. The task of implicitly performing complex computations in higher-dimensional spaces without explicitly calculating the coordinates of the data points in that space is taken up by kernels.

   Here we have used the "SVC()" function from scikit-learn with Radial Basis Function (RBF) kernel which is suitable for cases where there's no clear linear separation in the original feature space.

3. Random Forest Classifier
   Random Forest is a popular ensemble learning algorithm used for classification and regression tasks. It works by creating multiple decision trees during training and combining their outputs to make predictions. Each decision tree is trained on a random subset of the training data and a random subset of features. This randomness and diversity contribute to the algorithm's robustness and ability to handle complex datasets.

   In this study, we have used the "Random Forest Classifier ()" from Scikit-learn. It utilizes the Decision Tree Classifier() from the same library as the base estimator. We have put a 100 of these trees in the forest for efficient classification.

4. "XG Boost" Classifier
   "XG Boost" Classifier is an implementation of the Gradient Boosting algorithm that has been optimized for speed and performance. Gradient Boosting is a machine learning algorithm that works by combining the predictions of multiple weaker models (typically decision trees) to create a stronger, more accurate model. Gradient Boosting builds each new model to correct the errors made by the previous models, gradually improving its performance over iterations.

Here we employed “XG Boost Classifier()” from the “xg boost” library in python. We have used 100 estimators and 'objective' had been set to 'multi:softprob' as we have more than two classes in the target. The boosting learning rate ('eta') was set to 0.3.

5. “Hist Gradient Boosting” Classifier
   “Hist Gradient Boosting” Classifier is a machine learning classifier introduced in scikit-learn that belongs to the Gradient Boosting family. It's specifically designed to provide high-speed performance and scalability for large datasets by utilizing histogram-based techniques. This classifier is particularly effective when dealing with high-dimensional data. here we have set the 'learning rate' to 0.1 and the 'loss' as 'log_loss'.

6. Results

For ANN, “XGBoost”, “HistGradientBoosting”, Random Forest, and Naive Bayes classifiers sample size of five hundred thousand data points was employed. However, Support Vector Machine (SVM) could not process such a large dataset hence was limited to a subset of ten thousand data samples.

To ensure compatibility and robustness in training, we employed distinct scaling techniques based on the algorithms employed. For Artificial Neural Network (ANN), “XGBoost”, and “HistGradientBoosting” Classifier, the Mean Zero Unit Standard Deviation Scaling was applied, optimizing the convergence and performance of these models. In contrast, Max-Min Scaling was embraced for other models, as the former technique introduced unexpected NaN values that these models lacked the capacity to handle. The results produced are given in Table 1.

Table 1. Comparison of different models.

Model	Training Score	Test Score	Run Time
ANN	97.74%	97.31%	33.06s
XGBoost	98.55%	98.47%	7m 16s
HistGradientBoosting	98.09%.	98.19%	3m 54s
Random Forest	99.99%	98.73%	8m 3s
SVM (computed with only ten thousand data points.)	54.40%	54.18%	3m 18s
Naïve Bayes	80.74%	81.29%	3m 55s

Confusion Matrices:

1. ANN

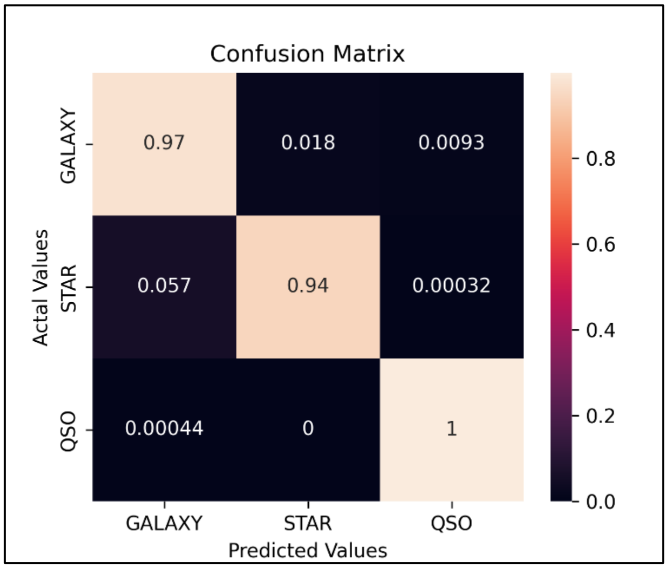


FIG 6. Confusion Matrix for ANN.

Our artificial neural network (ANN) demonstrates remarkable predictive capabilities, achieving nearly 100% accuracy in identifying QSOs and high accuracies of 97% for galaxies and 94% for stars. Nonetheless, there are minor instances where our model exhibits inaccuracies. Specifically, there is a 1.8% occurrence where galaxies are misclassified as stars, and a 5.7% occurrence where stars are incorrectly labeled as galaxies.

2. XGBoost

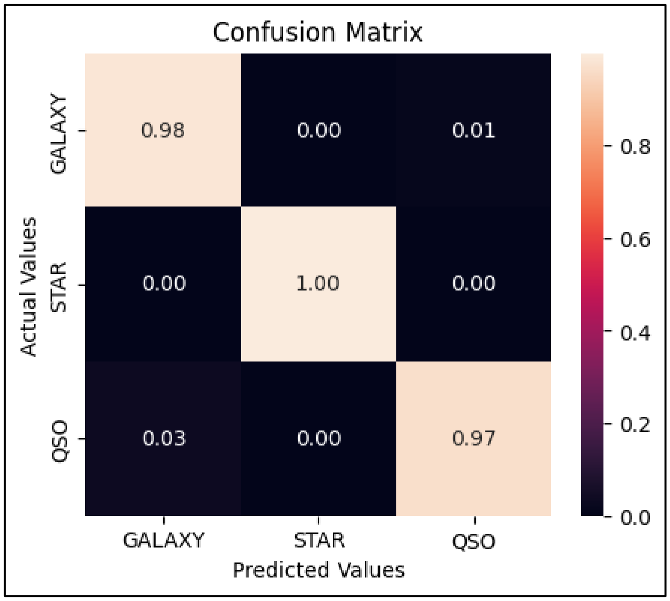


FIG 7. Confusion Matrix for XG Boost Classifier.

The XGBoost classifier does an excellent job in classifying stars with 100% accuracy. It identified QSOs and galaxies with acurracies of 97% and 98% respectively. It only misclassified 3% QSOs as galaxies and 1% galaxies as QSOs. Though the time required in the classification process is more than that of ANN.

3. HistGradientBoosting

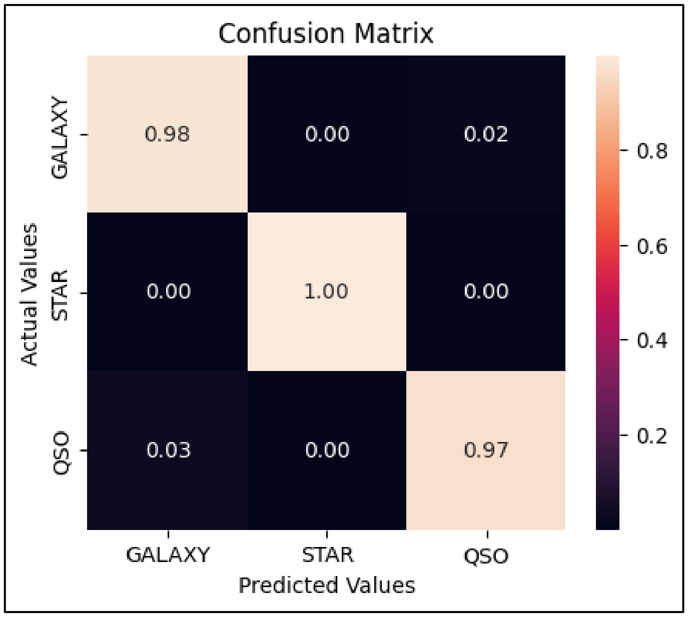


FIG 8. Confusion Matrix for Hist Gradient Boost Classifier.

The performance of HistGradientBoosting Classifier is close to that of XGBoost's. However, the time required by this algorithm was significantly less compared to XGBoost, without any compromise in the accuracy.

4. Random Forest

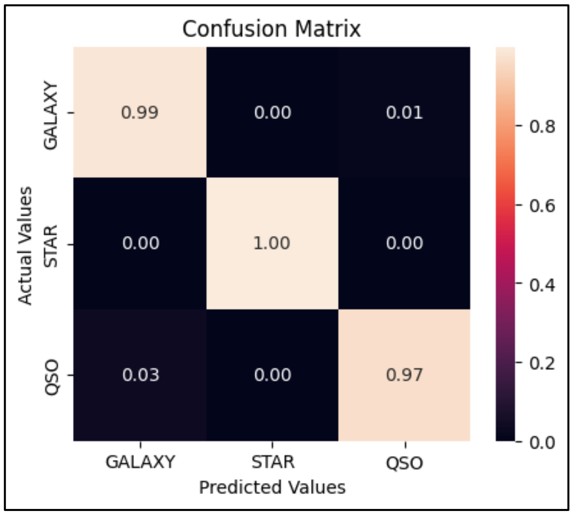


FIG 9. Confusion Matrix for Random Forest Classifier.

The Random Forest classifier gave the best performance, in terms of accuracy, among all the algorithms we have looked into. It correctly classicified 99% of galaxies, 100% of stars and 97% of QSOs. However, the time required is also significantly large compared to other models.

5. SVM

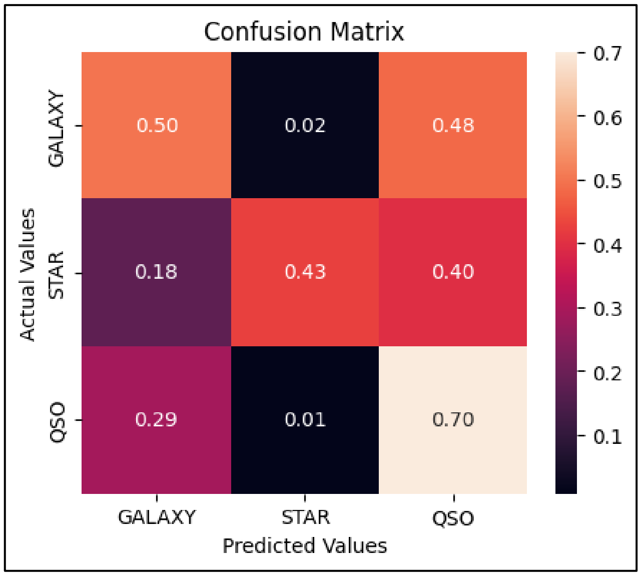


FIG 10. Confusion Matrix for Support Vector Machine Classifier.

Here we have tested the Support Vector Machine classifier with ten thousand data point. For higher number of data points, of the order of one lakh, it keeps on processing. The perfornce of this clasiifier is poorest among the models we tested in the study. For stars it falsely classifies more than half of the data as galaxies or QSOs. For galaxies the it predicted accurately for only 50% of the concerened data and 70% in the case of QSOs.

6. Naïve Bayes

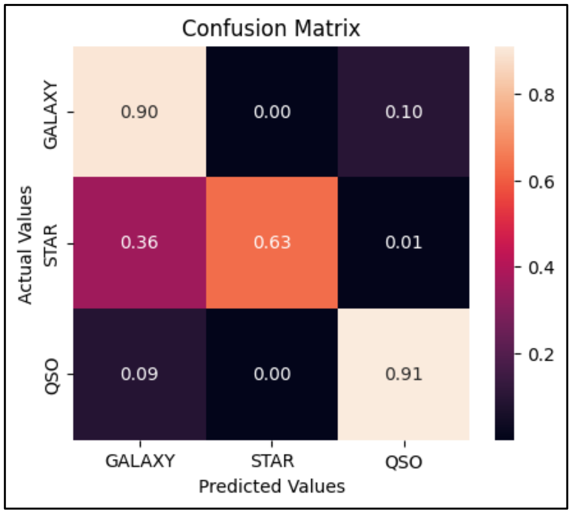


FIG 11. Confusion Matrix for Naïve Bayes Classifier.

The Naïve Bayes classifier did not perform so well as compared to the other models. It did accurately classify 91% of QSOs, 90% of galaxies and 63% stars but the misclassified values are significant too. About 36% of the stars have been inaccurately predicted as galaxies which is a tend we did not see in other models. The other models had the trend to misclaasify galaxies as QSOs. It completed the computation in a relatively small amount of time.

7. Conclusion

Among the algorithms explored, it was evident that Artificial Neural Networks (ANN) and ensemble classifiers, including Random Forest, XGBoost, and HistGradientBoosting, exhibited superior capabilities in effectively classifying these celestial objects.

Furthermore, our research highlighted an interesting observation concerning the Support Vector Machine (SVM) algorithm. While SVM is renowned for its versatility, we observed its limitations when confronted with vast amounts of data and numerous features, as has often been observed by others in literature[4]. It runs for over an hour without being able to fit the data, when it is provided with one lakh data points. The intricacy of finding a hyperplane to differentiate between classes became increasingly challenging in this scenario.

Also, Naive Bayes algorithm exhibited a decrease in its performance as the number of features increased. This outcome could be attributed to the underlying assumptions that features are conditionally independent given the class label.

Based on the confusion matrices, it can be deduced that the Artificial Neural Network exhibited superior performance in both accuracy and time efficiency. On the other hand, although the Random Forest classifier excelled in accuracy, it made a compromise in terms of processing time compared to the other models.

REFERENCES

[1] Mahalakshmi G S, Swadesh B, Aswin RRV, Sendhilkumar S, Swaminathan A and Surendran S, "Classification and Feature Prediction of Star, Galaxies, Quasars, and Galaxy Morphologies Using Machine Learning ", 2022.

[2] Wessam Salah Walid, "Stellar Classification - SDSS17 (4 ML Models)", available at https://www.kaggle.com/code/wessamwalid/stellar-classification-sdss17-4-ml-models/notebook, 2022.

[3] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, B.~Thirion, O.~Grisel, M.~Blondel, P.~Prettenhofer, R.~Weiss, V.~Dubourg, J.~Vanderplas, A.~Passos, D.~Cournapeau, M.~Brucher, M.~Perrot and E.~Duchesnay, "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, Vol.12 (pg 2825-2830), 2011.

[4] Janis Fehr, Karina Zapién Arreola and Hans Burkhardt , "Fast Support Vector Machine Classification of Very Large Datasets", *Data Analysis,* Machine Learning and Applications, pp 11–18, 2008.

[5] R. E. Uhrig, "Introduction to artificial neural networks", Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics, Orlando, FL, USA, pp. 33-37 vol.1, 1995.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pp. 35-39, 2019.

[7] Solan Digital Sky Survey, available at https://www.sdss.org/.

[8] Abdurro'uf et al., "The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data", The Astrophysical Journal Supplement Series, Volume 259, Issue 2, id.35, pp 39, 2022.