WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multi-Classification of Stars Based on Their Spectral Characteristics

***Nada M. Sallam*[1]**

[1]*Nile Higher Institute For Engineering And Technology , Al Manşūrah ,Egypt*
(Orcid ID: https://orcid.org/my-orcid?orcid=0000-0002-2921-7567)
**Email: nada.mohamed.sallam6@gmail.com*

ABSTRACT

As an emerging and pervasive topic, machine learning is making different development in different areas. It is also commonly used in the field of astronomy, where much research has been done utilizing machine learning for model prediction and data enhancement. In this study, for classifying stars in different types, I have been used one algorithm (Random Forest) to create predictive models. As a result of testing, the prediction accuracy reached about 98% of the random forest model, indicating the highest computational efficiency.

Keywords: *Spectral classification, machine learning, random forest, stars classification.*

1. Introduction

The Sloan Digital Sky Survey (SDSS) [1] achieved and put into operation with the development of science and technology, which helps getting data and information about different empyreal bodies. Also, machine learning technique has enabled different algorithms to train and efficiently classify large amounts of data gained from perception and collection [2] [3]. Academic research in this area is currently producing many results. When studying star and galaxy classification established on stacking ensemble learning, models are built using SVM, RF, and other models. The classification results are more accurate and robust compared to classical machine learning techniques. The Python language has been utilized in test process. It has characteristics such as simplicity [4], and the libraries included in it are enough. Classify galaxies, stars, and quasars into stars depend on decision trees, random forests, and support vector machine techniques, also compare the performance of different techniques with each other. This paper not only enhances the accuracy of the essential technique in star prediction, also gives a certain degree of accuracy in classifier model selection [5] [6].

2. Methodology

The dataset depends on 100,000 notices of space taken by the (SDSS). Each notices are characterized by feature columns and one class column that distinguishes between three types of stars [7].

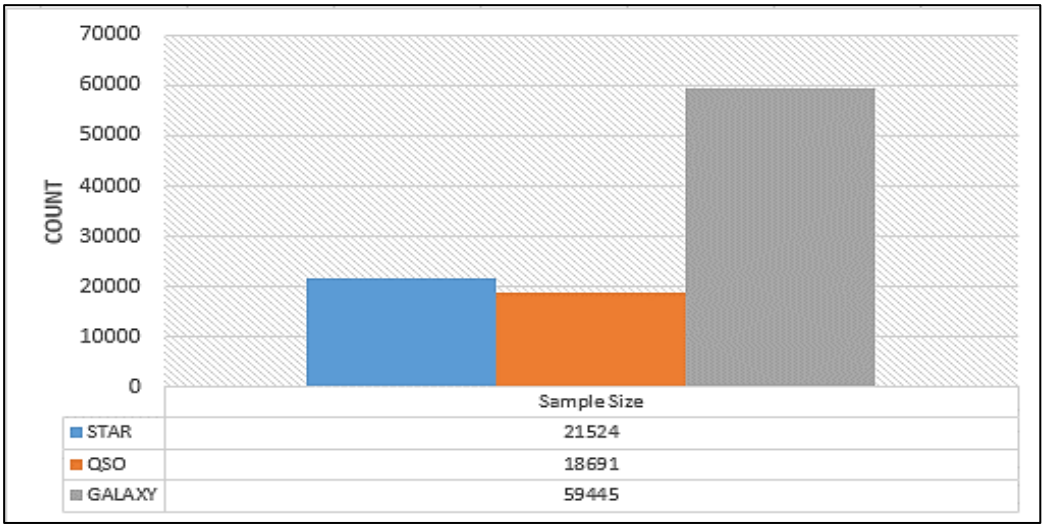

Figure 1. The size of patterns

obj_ID	alpha	delta	u	g	r
1.23766E+18	135.6891066	32.49463184	23.87882	22.2753	20.39501
1.23766E+18	144.8261006	31.27418489	24.77759	22.83188	22.58444
1.23766E+18	142.1887896	35.58244418	25.26307	22.66389	20.60976
1.23766E+18	338.7410378	-0.402827575	22.13682	23.77656	21.61162
1.23768E+18	345.2825932	21.1838656	19.43718	17.58028	16.49747
1.23768E+18	340.9951205	20.58947628	23.48827	23.33776	21.32195
1.23768E+18	23.23492643	11.41818762	21.46973	21.17624	20.92829
1.23768E+18	5.433176037	12.06518599	22.24979	22.02172	20.34126
1.23766E+18	200.2904754	47.19940232	24.40286	22.35669	20.61032
1.23767E+18	39.1496906	28.10284161	21.74669	20.03493	19.17553
1.23768E+18	328.0920762	18.22031048	25.77163	22.52042	20.63884
1.23766E+18	243.9866375	25.73828043	23.76761	23.79969	20.98318
1.23768E+18	345.8018744	32.67286785	23.17274	20.14496	19.41948
1.23768E+18	331.50203	10.03580205	20.8294	18.75091	17.51118
1.23766E+18	344.9847703	-0.352615781	23.20911	22.79291	22.08589
1.23766E+18	244.8245231	25.15456399	24.8868	22.13311	20.44728
1.23768E+18	353.2015224	3.080795936	24.5489	21.44267	20.95315
1.23768E+18	1.494388639	3.29174633	20.38562	20.40514	20.29996
1.23768E+18	14.38313522	3.214326196	21.82154	20.5573	19.94918
1.23765E+18	167.1316688	67.33993563	20.48292	18.67807	17.6168
1.23765E+18	171.9754246	67.74745014	22.13367	20.84772	18.96537
1.23766E+18	144.7852927	46.82649568	24.54793	22.33601	20.92259
1.23766E+18	145.2730374	46.96013381	25.44243	20.77028	19.6617
1.23766E+18	145.8830055	47.30048358	21.73992	21.53095	21.26763

Figure 2. samples from dataset

2.1. Data preprocessing

For different types of stars, the large differences in sample numbers prevent the training model from learning effectively, so leading to large deviations [8]. To overcome this challenge, oversample the data. Figure 3, depicted the samples for the stars that is up to 59445 [9].

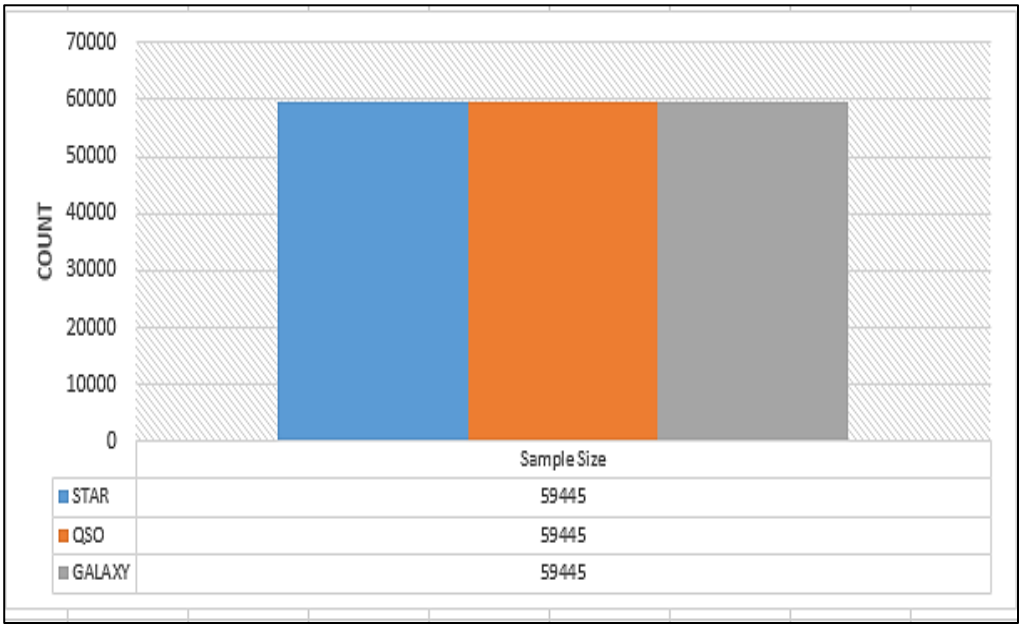

Figure 3. Each group after preprocessing

2.2. Classification Technique

The essential target of the model in manuscript is distinguishing between the three types of stars. Classification methods have been utilized in the suggested approach using RF classifier.

2.2.1. Random Forest (RF) Classifier

This technique is determined as a decision tree forest depending on algorithms loaded into random, separate trees. Best rated from multi decision trees and selected by majority. It is estimated one of the most important techniques. The essential challenge of this technique is overfitting is [10-14].

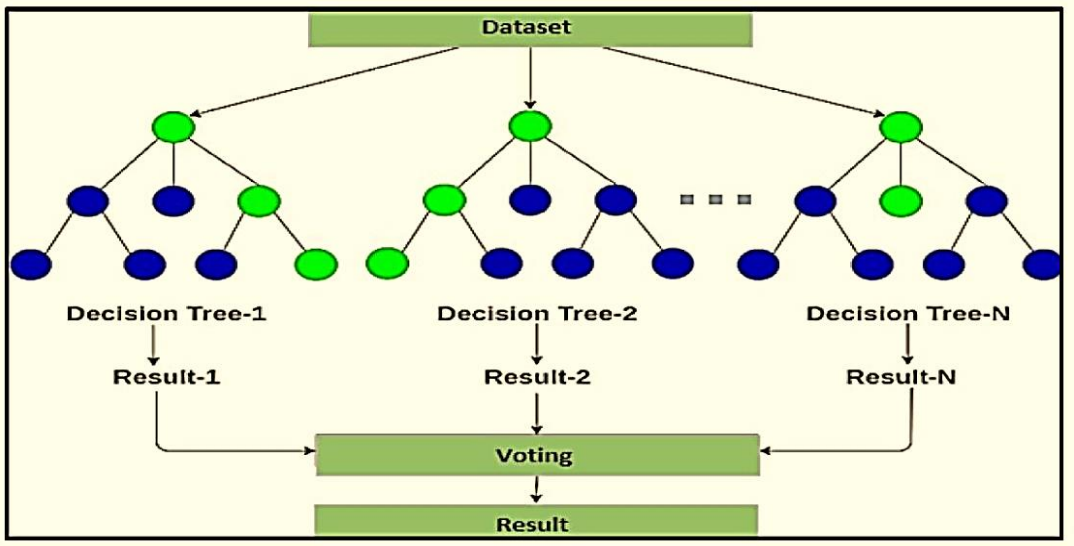

Figure 4. Hierarchy of random forest classifier

3. Results and discussion

The training outcomes is shown in figure 5 after using RF classifier. The accuracy of the RF model is 98%, and time for training requires 0.345 sec. The details of classification by the RF technique has been shown as confusion matrix in Figure 6.

	precision	recall	f1-score	support
GALAXY	0.98	0.99	0.98	11889
QSO	0.96	0.93	0.95	3792
STAR	0.99	1.00	1.00	4319
accuracy			0.98	20000
macro avg	0.98	0.97	0.97	20000
weighted avg	0.98	0.98	0.98	20000

Figure 5. The report for classification

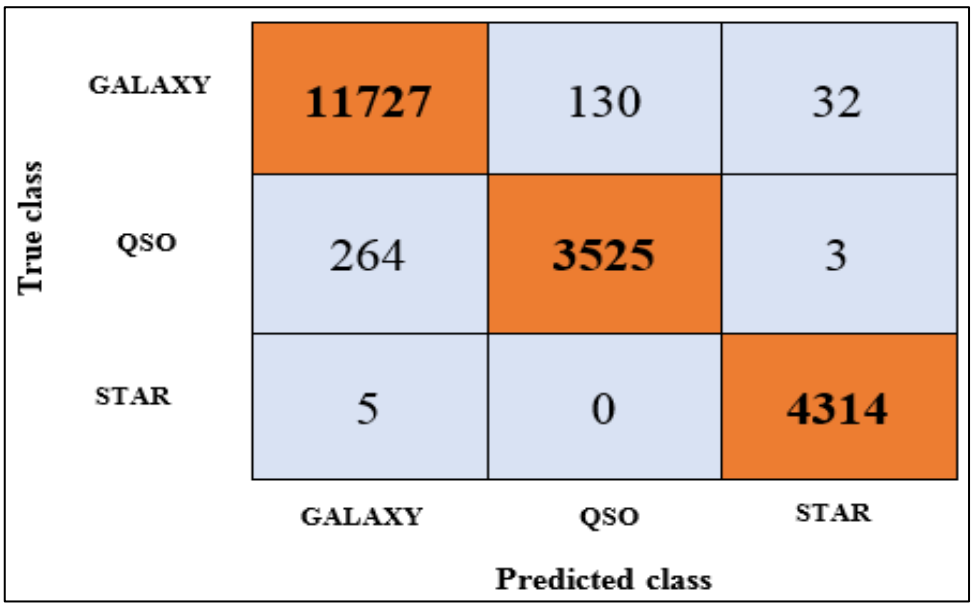

Figure 6. The confusion matrix after using RF classifier

4. Conclusion

This paper trained a technique using an RF classifier to classify all types of stars based on the spectral data gained from the Sloan Digital Sky Survey. From the training outcomes of the RF model, it can be shown that the RF model has the best execution in the dataset, that not only has the highest accuracy rate of 98%, but also has a high computing efficiency. Some challenges have been founded and needed to be enhanced in the training. As, in preprocessing stage for dataset, the aggregation of under and over-sampling may better to overcome over-fitting and enhance the training accuracy.

Data Availability

Data available on request from the authors

Conflict of Interest

I do not have any conflict of interest.

Funding Source

There is no funding sources

Authors' Contributions

Nada Mohamed Sallam is corresponded author, researched literature and conceived the study, put the protocol for manuscript, wrote the first draft of the manuscript. reviewed and edited the manuscript and approved the final version of the manuscript.

References

1) York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., ... & Yasuda, N. (2000). The sloan digital sky survey: Technical summary. The Astronomical Journal, 120(3), 1579.

2) Qi, Z. (2022). Stellar Classification by Machine Learning. In SHS Web of Conferences (Vol. 144, p. 03006). EDP Sciences.

3) Shifang, L. (2019). Application of Python in Artificial Intelligence [J]. Industrial Technology Innovation, 1(25), 33-34.

4) Zhang, H., Huang, T., Lv, Z., Liu, S., & Yang, H. (2019). MOOCRC: A highly accurate resource recommendation model for use in MOOC environments. Mobile Networks and Applications, 24, 34-46.

5) Zhou, Z. (2016). Learning Machine.

6) Zhou, Y., Liu, C. H., Wu, B., Yu, X., Cheng, G., Zhu, K., ... & Alfano, R. R. (2019). Optical biopsy identification and grading of gliomas using label-free visible resonance Raman spectroscopy. Journal of biomedical optics, 24(9), 095001-095001.

7) Accetta, K., Aerts, C., Aguirre, V. S., Ahumada, R., Ajgaonkar, N., Ak, N. F., ... & Kollmeier, J. A. (2022). The seventeenth data release of the Sloan Digital Sky Surveys: Complete release of MaNGA, MaStar, and APOGEE-2 data. The Astrophysical Journal Supplement Series, 259(2), 35.

8) Sallam, N. M., Saleh, A. I., Ali, H. A., & Abdelsalam, M. M. (2023). An efficient EGWO algorithm as feature selection for B-ALL diagnoses and its subtypes classification using peripheral blood smear images. Alexandria Engineering Journal, 68, 39-66.

9) Sallam, N. M., Saleh, A. I., Arafat Ali, H., & Abdelsalam, M. M. (2022). An Efficient Strategy for Blood Diseases Detection Based on Grey Wolf Optimization as Feature Selection and Machine Learning Techniques. Applied Sciences, 12(21), 10760.

10) Hötte, K., Tarannum, T., Verendel, V., & Bennett, L. (2022). Exploring Artificial Intelligence as a General Purpose Technology with Patent Data--A Systematic Comparison of Four Classification Approaches. arXiv preprint arXiv:2204.10304.

11) Sanlı, T., Sıcakyüz, Ç., & Yüregir, O. H. (2020). Comparison of the accuracy of classification algorithms on three data-sets in data mining: Example of 20 classes. International Journal of Engineering, Science and Technology, 12(3), 81-89.

12) Balaraman, S. (2020). Comparison of classification models for breast cancer identification using google colab.

13) Obaje, A., Onyancha, D., & Mirie, S. (2023). Pre-Treatment and Characterization of Cathode Active Material from Spent Lithium-IoN Batteries.

14) Ferrari, I. V., De Gregorio, A., Fuggetta, M. P., Ravagnan, G., Ali, W., Perrella, F., ... & Abdalla, M. (2023). Focus on Polydatin Interaction with Sirtuins Family: a Comparative Computational Analysis. Int. J. Sci. Res. in Biological Sciences Vol, 10(3).