Data-Driven Stellar Projection and Classification

Aman Kumar Sharma
M.Tech Scholar
School of Computer Science and Engineering
Vellore Institute of Technology,
Vellore – 632014, Tamil Nadu, India
email:amankumar.sharma2022@vitstudent.ac.in

Rajat Moundekar
M.Tech Scholar
School of Computer Science and Engineering
Vellore Institute of Technology,
Vellore – 632014, Tamil Nadu, India
email:rajat.moundekar2022@vitstudent.ac.in

Harsh Tyagi
M.Tech Scholar
School of Computer Science and Engineering
Vellore Institute of Technology,
Vellore – 632014, Tamil Nadu, India
email:harsh.tyagi2022@vitstudent.ac.in

Dr. Parthasarathy G
Assistant Professor
School of Computer Science and Engineering
Vellore Institute of Technology,
Vellore – 632014, Tamil Nadu, India
email:parthasarathy.g@vit.ac.in

Abstract*—Everything outside of Earth's atmosphere is the subject of astronomy. Stellar classification is used by astronomers to organize stars according to their spectral properties. Extra information on the elements, temperature, density, and magnetic field of the stars can be gleaned from their spectra. An essential aspect of astronomy is categorizing objects like galaxies, quasars, and stars. Classifying galaxies, stars, and quasars (luminous supermassive black holes) according to their spectral properties is the objective of this problem. In recent years, machine learning algorithms have demonstrated considerable potential for automating the categorization of stars. This study intends to assess the efficacy of several machine-learning algorithms for categorizing stars according to their physical attributes. The research will include gathering and pre-processing data from astronomy catalogs and surveys. The data will be divided into training and testing sets, and then a number of machine-learning techniques will be employed. The performance of these algorithms will be measured using measures like precision, recall, and accuracy. This project's outcomes will shed light on the efficacy of various ML algorithms for star classification and determine the optimal methods for this purpose. This work will benefit astronomers and astrophysicists whose research relies on precise star classifications. In addition, it will demonstrate the capacity of machine learning to automate and accelerate the star categorization process.

***Keywords*— Spectral Type, Spectral Classification, Machine learning, Stellar Classification, Astronomy.**

I. Introduction

Numerous sky survey technology projects, such as the Sloan Digital Sky Survey (SDSS), have been completed and put into service[1] as a consequence of advancements in science and technology, making it simpler to acquire information about many celestial bodies. At the same time, a variety of ML algorithms may be effectively trained on vast volumes of data gathered through observation and gathering.[2]. It is critical to investigate numerous stars' properties and broaden our understanding of the universe. Academic research in this field is now making tremendous progress. The Python programming language will be utilized for this project, which provides simplicity, extensibility, and other features[3]. The provided library is sufficient as well. KNN, Decision Tree, Logistic Regression, Random Forest, XG Boost Classifier, and an ensemble model combining all of these are used to categorize galaxies, stars, and quasars in stars, and their relative performances are compared. The hyperparameters are also fine-tuned to improve accuracy. This study can potentially improve efficiency and accuracy in star classification over the original method and tell us about the usefulness of different training models when applied to different objects by analyzing the results of our classifications.

II. Literature Survey

In Stellar Classification by Machine Learning by Zhuliang Qipaper, the author provides a literature review of recent research on stellar classification using machine learning techniques. The paper goes on to review several machine-learning techniques used for stellar classification, including ANNs, SVM, decision trees, and random forests. The author discusses the strengths and weaknesses of each technique and provides examples of their applications in stellar classification. Challenges in this field include the need for large and diverse datasets and the selection of appropriate features to represent the spectra of stars [4].

In Stellar and Pulsar Classification using ML by Jishant Talwar et al, the authors present an approach for classifying stars and pulsars using machine learning techniques. They apply several ML techniques, including SVM, Decision Trees, and Random Forests, to classify the stars and pulsars based on these features. Challenges in this include the need for large and diverse datasets and the selection of appropriate features to represent the data. Additionally, the authors faced the challenge of dealing with imbalanced datasets, where the number of examples in each class is significantly different. This can lead to biased classification results if not properly addressed [5].

Stellar Classification with Fuzzy Clustering and Self-Organizing Maps by S. Saha et al.: This paper proposes a fuzzy clustering and self-organizing map-based approach to classifying stars based on their spectra. One challenge the authors faced was the need to choose appropriate parameters for their algorithm to achieve accurate classification. Additionally, the authors had to select appropriate features to represent the spectra in a way that the clustering algorithm and self-organizing map could effectively classify [6].

Stellar spectral classification and feature evaluation based on a random forest by Xiang-Ru Li the authors present an approach for the spectral classification of stars using a random forest algorithm. They apply a random forest algorithm to classify the stars based on these features.

Challenges in this include the need for large and diverse datasets and the selection of appropriate features to represent the spectra. The authors also faced the challenge of dealing with imbalanced datasets, where the number of examples in each class is significantly different [7].

In stellar spectral classification using PCA and ANNs by Harinder P. Singh et al, the authors present an approach for spectral classification of stars using PCA and ANNs. Their approach involves reducing the dimensionality of the spectral data and then applying ANNs for classification. Challenges in this include the need for large and diverse datasets and the selection of appropriate features to represent the spectra. Additionally, selecting appropriate hyperparameters for the ANN model can be challenging, and there is a potential loss of information when reducing the dimensionality of spectra using PCA [8].

III. Proposed Model

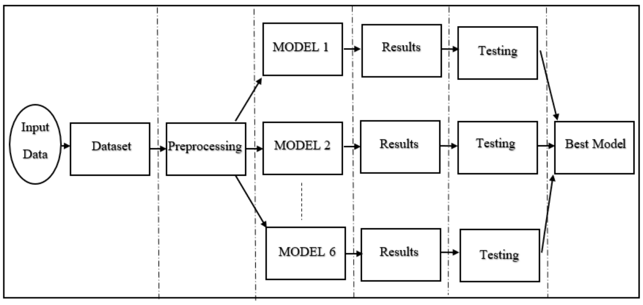

Fig.1. Proposed Model

A. Dataset Overview

The data consists of 100,000 space observations made by SDSS. Each data point has 17 feature columns that characterize it, and one class column that designates whether it is a star, galaxy, or quasar [9].

B. Data Preprocessing

1) Distribution of the data

The data's class distribution and class proportion are displayed in the charts below.

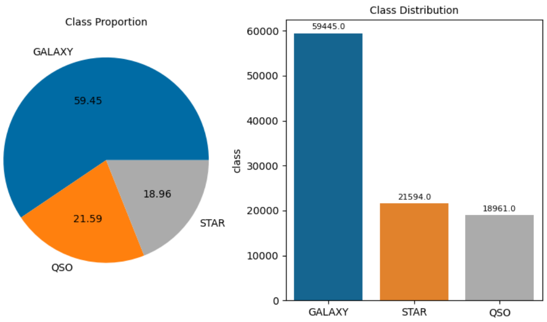

Fig.2. Class distribution and proportion

2) Multivariate analysis considers multiple features to analyze the data.

Pairplot:

Pairwise relationships in the data can be seen in the pairplot [10]. Pairwise relationships between features are depicted in the following pairplot. In summary, we see that u, g, r, i, and z are all positively connected.

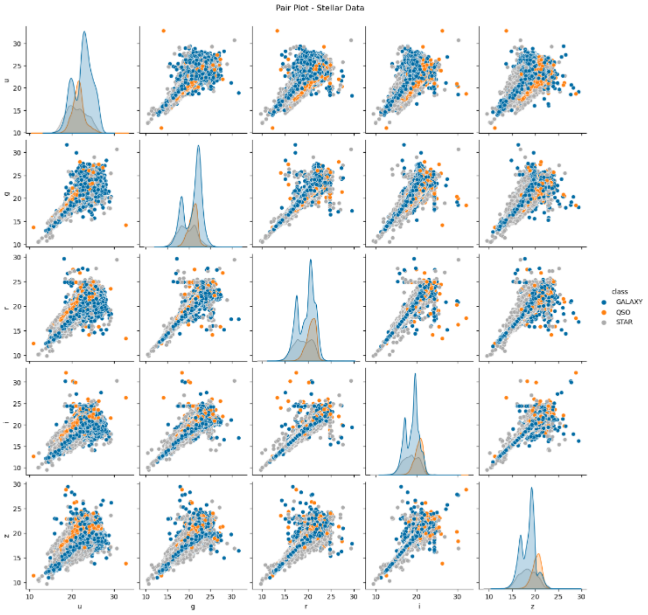

Fig.3. Representation of Pairplots for u, g, r, i, and z features

3) A data standardization technique transforms many data sets into a uniform framework. After data has been acquired from various sources, it must be transformed before being put into target systems [11].

The range of features in our dataset is different so we have done data standardization

C. Model Introduction

1) K Nearest Neighbour: KNN is a well-liked ML technique for both classification and regression. Predictions are made using the input data's similarity to its k nearest neighbors, which is a non-parametric technique [12]. We have three formulas to calculate the distance metric in KNN but here we are using the Minkowski Distance. The formula for Minkowski distance:

$$D\left(X,Y\right)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{\frac{1}{p}} \tag{1}$$

2) Decision Tree(D.T.): It is a typical machine-learning technique used for classifying data. A decision tree is a graphical representation of a classification tree, complete with a root node, branch nodes, and leaf nodes. Each leaf node represents a different prediction, and the data inside each node is distributed to its children based on the outcomes of attribute tests. The information entropy is used to choose the optimal splitting method while training a decision tree model [4]. The entropy of information is defined as follows:

$$Ent(D)=-\sum_{k=1}^{|y|}p_k log_2 p_k \tag{2}$$

3) Random Forest(R.F.): A random forest is a more complex variant of bagging. It creates Bagging on the basis of the D.T. and includes a quick selection of attributes. To rephrase, the D.T. weighs measures like information entropy and gain rate in determining which feature is superior. Random Forest picks a subset of K characteristics at random from the attribute set of every node in the initial D.T.[4].

4) XG Boost: Extreme Gradient Boosting, or XGBoost, is a well-liked machine learning method that excels in a variety of tasks like classification, regression, and ranking. The gradient boosting paradigm is being used in practice to create strong ensemble models from a variety of very feeble prediction models, often decision trees [13].

5) Logistic Regression: Logistic regression is a well-liked statistical model for binary classification issues, in which the goal is to foretell the probability that an instance will belong to a given class. Contrary to what its name suggests, logistic regression is a classification algorithm. It uses the logistic function to show how the binary dependent variable and one or more continuous or categorical independent variables are related [14].

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

6) Ensemble Voting Classifier Model: The ensemble voting classifier model is an ML technique that combines multiple individual classifiers (base models) to make predictions. It leverages the wisdom of the crowd by aggregating the predictions from each base model and selecting the class label that receives the majority of the votes[15]. Here we combined Logistic regression, random forest, and XG-Boost to make an ensemble voting classifier model.

D. Hyperparameter tuning: Hyperparameters control the learning process of the model. In this process, the learning model is exposed to a set of values for each parameter, and the best parameter value is selected. This process is called tuning. The performance metric for tuning is typically measured by cross-validation on the training set. We used GridSearchCV for tuning the hyperparameters [10].

IV. Results and discussion

Table 1. Classification Report for Various Models

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
KNN	0.92	0.93	0.89	0.91
D.T.	0.97	0.96	0.96	0.96
R.F.	0.98	0.98	0.97	0.97
XG Boost	0.98	0.98	0.97	0.97
Logistic Regression	0.96	0.96	0.95	0.96
Ensemble Model	0.98	0.98	0.98	0.98
Qi.Z.[4]	0.97	0.97	0.97	0.97

The Ensemble Voting Classifier model is shown to have an overall accuracy of 98%. When compared to other models, the Ensemble Voting Classifier Model has the highest accuracy, precision, recall, and F1 score. and when we refer to our base paper the overall accuracy is 97% for the support vector machines model which is low, and the model computational time is the longest which is improved in our new ensemble voting classifier method which has higher accuracy and low training time.

V. Conclusion

In this work, astronomical objects including galaxies, stars, and quasars were categorized using ML techniques like KNN, decision tree, Logistic Regression, Random Forest, XG-Boost, and Ensemble Model, all based on data from the SDSS. The Classification Report of the six models shows that the Ensemble Voting Classifier Model performed best and achieved an accuracy rate of 98% while using very few processing resources. Furthermore, all six algorithms achieved outstanding results when tasked with classifying galaxies, stars, and quasars.

References

[1] YORK D G, ADELMAN J, ANDERSON JR J E, et al. The Sloan digital sky survey: technical summary[J]. The Astronomical Journal, 2000, 120(3); 1579

[2] Zhou Jie, Zhu Jianwen. Research on machine learning classification problem and algorithm [J] Software, 2019, 40 (7): 205-208

[3] Liu Shifang. Application of Python in Artificial Intelligence. Industrial Technology Innovation, 2019, 25

[4] Qi, Z. (2022). Stellar Classification by Machine Learning. SHS Web of Conferences. https://doi.org/10.1051/shsconf/202214403006

[5] Jishant Talwar et al, "Stellar and Pulsar Classification using Machine Learning" HANS SHODH SUDHA, Vol. 1, Issue 4, (2021), pp. 71-80

[6] Saini, Naveen & Chourasia, Shubham & Saha, Sriparna & Bhattacharyya, Pushpak. (2017). A Self Organizing Map Based Multi-objective Framework for Automatic Evolution of Clusters. 672-682.

[7] Li, Xiangru & Lin, Yang-Tao & Qiu, Kai-Bin. (2019). Stellar spectral classification and feature evaluation based on a random forest. Research in Astronomy and Astrophysics. 19. 111. 10.1088/1674-4527/19/8/111.

[8] Singh, Harinder & Gulati, Ravi & Gupta, Ranjan. (2002). Stellar Spectral Classification using Principal Component Analysis and Artificial Neural Networks. Monthly Notices of the Royal Astronomical Society. 295. 312 - 318. 10.1046/j.1365-8711.1998.01255.x.

[9] Abdurro'uf et al., The Seventeenth data release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 DATA (Abdurro'uf et al. submitted to ApJS) [arXiv:2112.02026].

[10] https://towardsdatascience.com/stellar-classification-a-machine-learning-approach-5e23eb5cadb1 last accessed on 10.05.2023

[11] https://www.simplilearn.com/what-is-data-standardization-article#:~:text=Data%20standardization%20is%20converting%20data,YYYY%2DMM%2DDD). Last accessed on 10.05.2023

[12] https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html Last accessed on 12.05.2023

[13] https://www.geeksforgeeks.org/xgboost/ Last accessed on 12.05.2023

[14] https://towardsdatascience.com/logistic-regression-explained-from-scratch-visually-mathematically-and-programmatically-eb83520fdf9a Last accessed on 13.05.2023

[15] https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html Last accessed on 14.05.2023

[16] https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall Last accessed on 18.05.2023

[17] https://en.wikipedia.org/wiki/Celestial_sphere Last accessed on 18.05.2023

[18] https://en.wikipedia.org/wiki/Celestial_equator Last accessed on 19.05.2023

[19] https://www.esa.int/Science_Exploration/Space_Science/What_is_red_shift Last accessed on 19.05.2023

[20] https://en.wikipedia.org/wiki/UBV_photometric_system Last accessed on 20.05.2023