AstroClassify: A ML Framework for Celestial Object Classification

Deekshith Reddy Yeruva
SCAI, Arizona State University
Tempe, AZ
dyeruva@asu.edu

Saibhaskara Durga Mani Surya Dheeraj Kanuri
SCAI, Arizona State University
Tempe, AZ
sskanuri@asu.edu

Sidhartha Reddy Gundarapu
SCAI, Arizona State University
Tempe, AZ
sgundara@asu.edu

Surya Raman Nagarajan
SCAI, Arizona State University
Tempe, AZ
snagar40@asu.edu

Yaswanth Kumar Reddy Karri
SCAI, Arizona State University
Tempe, AZ
ykarri@asu.edu

***Abstract*- In the realm of astronomy, accurate classification of celestial objects is pivotal for understanding the universe's intricacies. Our research utilizes the extensive Sloan Digital Sky Survey DR17 dataset [10] to classify stars, galaxies, and quasars according to their spectral properties utilizing advanced machine learning algorithms. The task involves formulating a multi-class classification problem, assigning each observation a distinct class, with the overarching objective of developing a robust categorization model. By meticulously addressing the challenges of feature selection, class imbalance, and model optimization, we aim to enhance classification accuracy and computational efficiency. Through rigorous data analysis and iterative model refinement, this project strives to unlock new avenues for astronomical research, contributing to the ongoing quest for deeper insights into the cosmos and its myriad phenomena.**

I. INTRODUCTION

A. Background

Stellar classification, a fundamental aspect of astronomy, involves categorizing celestial objects into stars, galaxies, and quasars classes in light of their spectral characteristics . Manual classification methods are time-consuming, prompting the need for automated solutions.

B. Problem

The task of classifying stars, galaxies, and quasars presents challenges due to the dataset's complexity, including feature correlation and imbalanced class distributions. The manual classification of celestial objects can be time-consuming, taking minutes to hours. Machine Learning (ML) models offer a solution, leveraging multi-wavelength data to improve classification accuracy and significantly reduce classification time for astronomers. Additionally, the classification task encounters several challenges, such as dealing with heterogeneous data sources where observations from different telescopes or instruments introduce variations in data quality and distribution, necessitating harmonization for accurate classification. Furthermore, temporal variability in spectral characteristics due to cosmic events or observational conditions adds complexity, requiring careful consideration to account for

such fluctuations during classification. The dataset's high dimensionality poses challenges in feature selection and model training, while sparse data for certain features makes deriving meaningful insights difficult, particularly for rare celestial objects. Additionally, the complex, non-linear relationships inherent in spectral data necessitate the use of advanced ML algorithms capable of capturing intricate patterns. Effective classification also requires domain knowledge in astronomy to interpret spectral features and understand their significance in differentiating between classes.

C. Importance

The importance of this project lies in its potential to revolutionize the way astronomers classify celestial objects. By reducing the classification time and improving the classification accuracy, this project can contribute to a better understanding of the universe. Accurate classification of celestial objects facilitates deeper insights into the universe's composition and evolution. ML techniques offer a promising avenue for efficient and precise categorization, enabling astronomers to expedite their research endeavors.

D. Existing Literature

Several studies in the field of astronomy have extensively explored the application of ML algorithms for stellar classification. For instance, Yang and Zhou et al.[1] utilized support vector machines (SVM) and convolutional neural networks (CNN) to classify stars and galaxies, demonstrating the efficacy of ML techniques in astronomical classification tasks.

Similarly, Sánchez and Cuadros et al. [2] investigated the use of random forest and gradient boosting algorithms for classifying celestial items according to photometric and spectroscopic data, highlighting the significance of ensemble methods in improving classification accuracy.

Furthermore, research by Lambert et al. [3] focused on feature engineering techniques such as principal component analysis (PCA) and manifold learning to extract informative features from multi-wavelength data, enhancing the discriminatory power of classification models. These studies underscore the importance of robust methodologies encompassing feature selection, data preprocessing, and model evaluation in achieving accurate and efficient classification of celestial objects..

E. System Overview

The system encompasses several key components and processes essential for accomplishment of the ML system. Beginning with data preprocessing, the system undergoes cleaning and standardization to ensure data consistency and uniformity. Feature engineering follows, involving the extraction of relevant features and the creation of new variables to enhance predictive power. Model selection plays a pivotal role, where various ML algorithms like random forests, and gradient boosting, naïve baye’s are explored to identify the most suitable model for the classification task.

Subsequently, models are trained on preprocessed data using cross-validation techniques to optimize performance and prevent overfitting. Model evaluation assesses performance using metrics like accuracy and precision, facilitating the selection of the best-performing model for deployment. Lastly, hyperparameter tuning fine-tunes model parameters to further enhance performance and effectiveness. This comprehensive overview

underscores the intricate processes involved in developing and deploying a robust ML system for celestial object classification.

F. Data collection

The Sloan Digital Sky Survey DR17 dataset, consisting of 100,000 observations with 17 features, serves as the foundation for our classification task.

G. Components of the ML system

The ML system comprises several essential components aimed at effectively classifying celestial objects based on their spectral characteristics. Firstly, data preprocessing has a vital role in ensuring the dataset's quality and consistency. This involves handling outliers, addressing imbalanced class distributions through techniques like Synthetic Minority Oversampling Technique (SMOTE), and standardizing features to maintain uniformity and prevent bias during model training. Feature selection follows, where relevant features are identified while redundant ones are removed to improve model performance and computational efficiency. Next, model selection involves experimenting with various ML algorithms such as Random Forests, and Gradient Boosting to identify the most suitable ones for stellar classification. Lastly, experimental results are assessed using parameters like F1-score and accuracy., revealing details about the performance of the model and guiding further iterations and refinements. Through meticulous implementation and evaluation of these components, our ML system aims to develop a robust classification model for celestial objects, advancing our understanding of the universe's vast complexities.

II. IMPORTANT DEFINITIONS

A. Data

Each observation corresponds to a celestial object captured by the survey. The dataset includes 18 attributes, with 10 of type float, 7 of type int, and 1 of class object type which mentions the type of the celestial object (star, galaxy, or quasar).

The dataset provides a comprehensive snapshot of celestial phenomena across multiple wavelengths, ranging from ultraviolet to infrared, enabling us to capture diverse spectral characteristics. Furthermore, the dataset includes metadata such as object identifiers, right ascension, declination, and redshift values, facilitating detailed analysis and classification. Additionally, the dataset is supplemented with plate and fiber IDs, indicating the observational parameters and instrumentation used in data collection. This rich and extensive dataset offers invaluable insights into the universe's composition and behavior, serving as a cornerstone for our ML-based classification approach.

B. Prediction Target

In this context, the prediction target is the "class" attribute in the dataset, which identifies the type of celestial object (e.g., star, galaxy, quasar).

C. Variables/concepts in the data

1) obj_ID: Each celestial object in the image catalog that the CAS (Catalog Access Services) uses is given a distinctive value in the Sloan Digital Sky Survey (SDSS) [10].

2) Alpha: Right Ascension angle (α) represents the east-west coordinate in the equatorial coordinate system, measured in hours, minutes, and seconds [10].
3) Delta: Declination angle (δ) represents the north-south coordinate in the equatorial coordinate system, measured in degrees, minutes, and seconds [10].
4) r, u, g, z, i: These are the system filters used in the SDSS dataset, representing ultraviolet, green, red, near-infrared, and infrared wavelengths, respectively [10].
5) run_ID: Run Number is a unique identifier used to identify the specific scan during the SDSS observation process [10].
6) rereun_ID: Rerun Number is a supplementary identifier specifying how the image was processed during the observation process [10].
7) cam_col: This represents the scanline within the run, identifying the specific camera used during observation [10].
8) field_ID: This identifies each field within the observation process [10].
9) spec_obj_ID: This is the ID used for optical spectroscopic objects, ensuring that different observations sharing the same spec_obj_ID belong to the same class [10].
10) class: This represents the classification of celestial objects into categories such as stars, galaxies, and quasars [10].
11) redshift: This value represents the expansion of the light's wavelength due to the Doppler effect, providing information about the distance and velocity of celestial objects [10].
12) plate: This recognizes each plate used in the SDSS observation process [10].
13) MJD (Modified Julian Date): MJD is used to specify the time at which a specific piece of SDSS data was collected, providing a timestamp for observations [10].
14) fiber_ID: This is used to identify the fiber that is directed at the illumination at the observation's focal plane, allowing for spectroscopic analysis of celestial objects [10].

D. Problem Statement

The classification of astronomical objects according to their spectral properties presents a multifaceted challenge in the field of astronomy. The aim of our project is to develop and deploy a robust data mining solution that can accurately classify celestial objects into specific categories (stars, galaxies, and quasars) using ML techniques applied to observational data. Manual classification methods are time-consuming and often subjective, whereas ML models offer the potential for efficient and objective categorization, significantly reducing classification time for astronomers. However, the task faces several challenges, including handling imbalanced class distributions, navigating the complex domain of astronomy, and optimizing models for accurate classification while ensuring computational efficiency. By addressing these challenges and leveraging the rich dataset acquired from SDSS , our project aims to contribute to a deeper understanding of the universe's composition and behavior through automated and precise classification of celestial objects based on their spectral properties.

III. Overview Of The Proposed Approach

The proposed approach for the data mining model involves several key steps to analyze and classify astronomical objects. Here's the overview of the proposed approach:

1) Data Preprocessing and Outlier Detection: The initial step involves importing the necessary libraries and loading the dataset. Following this, the data is inspected to understand its structure and distribution. Outlier detection methods are applied to find and eliminate anomalous data points from the dataset. This guarantess the quality and integrity of the data utilized in the analysis [4].
2) Feature Selection and Correlation Analysis: After preprocessing, feature selection methods are employed to choose related attributes for the classification task.
3) Handling Class Imbalance: Our dataset exhibits class imbalance, where certain classes are underrepresented. Imbalance mitigating techniques are used to generate synthetic samples for minority classes [5]
4) Normalization and Model Training: Before model training, the features are standardized to ensure all variables have the same scale.
5) Model Evaluation and Selection: When evaluating trained models, we used various metrics such as accuracy, and F1-score. Confusion matrices are utilized to assess model performance and identify any misclassifications.Additionally, Hyperparameter tuning techniques are employed to optimize model parameters and improve classification accuracy.
6) Ensemble Learning and Stacking: Ensemble learning methods are explored to combine the predictions of multiple base classifiers into a single meta-classifier. This approach leverages the strengths of different algorithms to enhance predictive performance and generalize well on unseen data [7].
7) Results and Conclusion: The final step involves summarizing the performance of each model and the ensemble method. The approach with the highest accuracy and robustness is recommended for classifying astronomical objects. Insights from the analysis can provide valuable information for astronomers and researchers in understanding celestial phenomena and exploring the universe further.

By following this systematic approach, the data mining model can effectively classify astronomical objects and contribute to advancements in astronomy and astrophysics research.

IV. Technical Details of the Proposed Approach

1) Data Preprocessing and Outlier Detection: The data preprocessing phase involves several steps to prepare the dataset for modeling. Initially, Local Outlier Factor (LOF) is applied to identify outliers in the dataset. LOF allocates an outlier score to every data item, and those with scores below a threshold (-1.5 in this case) are considered outliers and subsequently removed from the dataset. This step ensures the robustness of the models by excluding potentially noisy data points [4].
2) Feature Engineering and Transformation: Following outlier removal, feature correlation analysis is performed using a heatmap visualization technique. Understanding the connections between various attributes and the target variable is made easier by this analysis. Highly correlated features can have an

impact on the performance of ML models, so redundant features are identified and removed from the dataset. Additionally, categorical variables are encoded into numerical format to facilitate model training [.

3) Handling Class Imbalance: Class imbalance is addressed utilizing the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples for the class with lesser dat items to balance the class distribution. This technique helps prevent bias towards the majority class during model training, resulting in higher precise predictions. After applying SMOTE, dataset is resampled to achieve a balanced class distribution [5].
4) Model Training and Evaluation: Multiple classification algorithms are trained on the preprocessed data, including Bernoulli Naive Bayes, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Gradient Boosting. These models are accessed leveraging various performance metrics such as accuracy score, confusion matrix, and classification report. Yellowbrick visualization tools are utilized to visualize the performance metrics, providing insights into model performance and areas for improvement [6].
5) Ensemble Methods: Ensemble methods such as stacking and voting are explored to further improve predictive performance. Stacking uses a meta-learner to integrate the predictions of several base models, while voting aggregates the predictions of individual models to make final predictions. The ensemble models are evaluated using cross-validation to ensure generalizability and robustness. Additionally, a blending ensemble approach is also employed to combine predictions from base models, further enhancing predictive performance [7].
6) Hyperparameter Tuning: Hyperparameter tuning is performed using Randomized Search Cross-Validation to find the optimal hyperparameters for Random Forest, KNN, Logistic Regression, and Gradient Boosting models. This step helps optimize model performance by fine-tuning model parameters. The best combination of hyperparameters is then used to train the models, resulting in improved predictive accuracy [9].
7) Model Evaluation with Best Parameters: Finally, using the test dataset, the tuned models' performance is assessed. The accuracy, confusion matrix, and classification report are computed to assess the models predictive power and generalization ability.

By following these steps, the proposed data mining model effectively addresses data preprocessing, feature engineering, class imbalance, model training, evaluation, and hyperparameter tuning, resulting in robust and accurate predictive models for star classification.

V. EXPERIMENTS

E. Initial data exploration:

Through the initial data exploration the dataset provides information about celestial objects categorized into three classes: GALAXY, STAR, and QSO (Quasi-Stellar Object). Each class has a corresponding count of instances as shown the image given

below

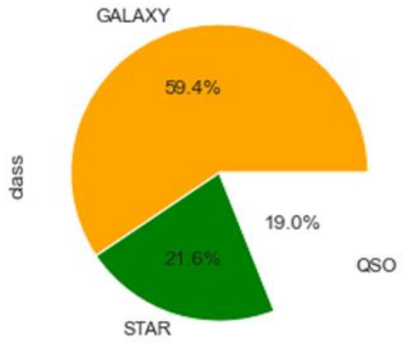

Fig 1: Value counts of each class

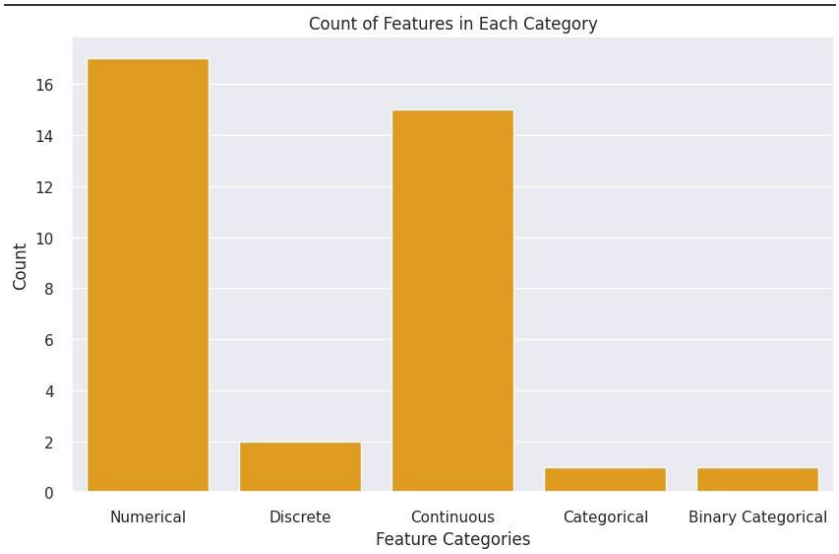

Fig 2: Visualization to show the feature types and their counts

F. Baseline

a. Baseline data preprocessing: In the initial data preprocessing, several key steps were taken to ensure data quality and enhance model performance. Outliers were addressed using the interquartile range method, minimizing their impact on analysis. Feature selection based on correlation analysis helped prioritize relevant attributes by removing those with low correlation to the target variable. Imbalanced data issues were tackled through SMOTE oversampling, promoting a more balanced representation of classes. Standardization maintained feature uniformity, while duplicate entries were eliminated to preserve dataset integrity. Finally, label encoding converted

the target variable into numeric format, facilitating model compatibility. These preprocessing steps collectively contributed to improving the dataset's suitability for accurate and robust data analysis and modeling tasks.

b. Baseline Model performances: Following the preliminary data preprocessing and feature selection steps, a variety of ML models, including KNN Classifier, Logistic Regression, Gaussian Naive Bayes ,Bernoulli Naive Bayes, Random Forest, and Gradient Boosting, were trained and evaluated to assess their predictive accuracies. The performance of each model was tested to determine their effectiveness in handling the dataset. Below are the accuracies of the baseline models.

```
Accuracy Scores:
GaussianNB:          0.6300549317965678
BernoulliNB:         0.7731189053384622
KNeighbours:         0.7971214035286937
LogisticReg:         0.6302678457367532
RandomForest:        0.9205132572531652
GradientBoosting:    0.9445296088116732
```

Fig 3: Accuracy scores of models

Confusion matrices for the two models that achieved the highest accuracy are given below

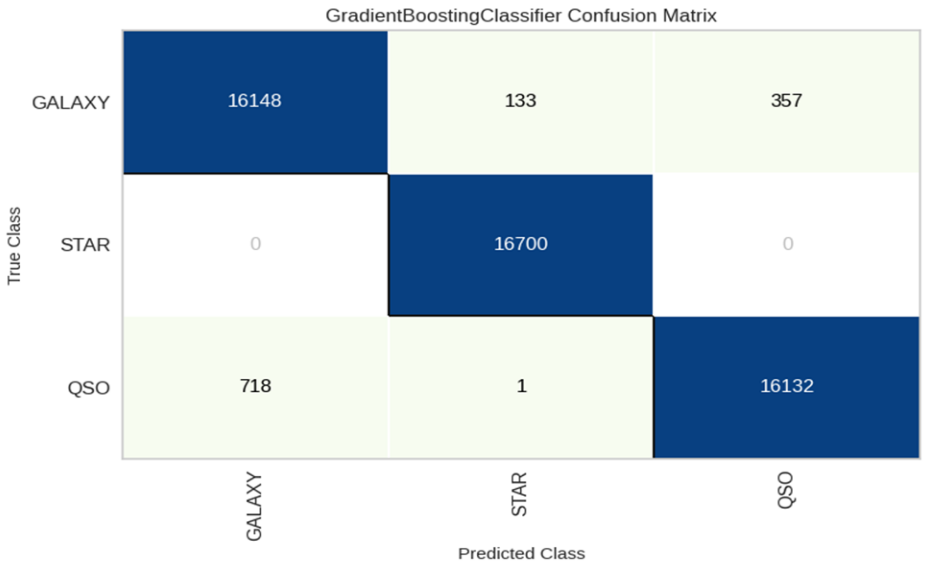

Fig 4: Confusion matrix for Gradient Boosting Classifier

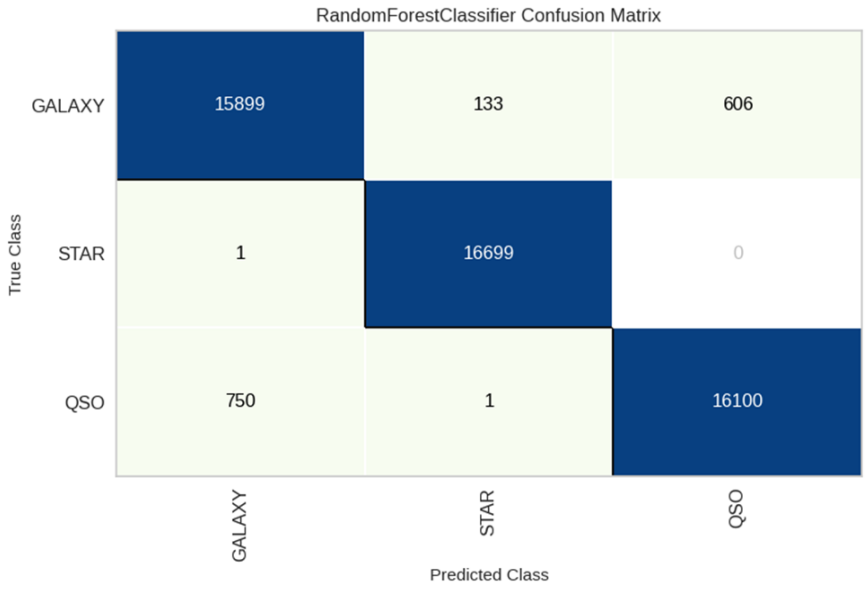

Fig 5: Confusion matrix for Random Forest Classifier

G. Improvements

a. Improvements in data preprocessing: In refining the data preprocessing pipeline, several enhancements were implemented. Firstly, the Local Outlier Factor (LOF) algorithm replaced the Interquartile Range (IQR) method for outlier detection and removal, offering a more nuanced approach to identifying anomalies within the dataset. Lastly, feature standardization was performed using StandardScaler, ensuring uniform scale across numerical features, thereby preventing any feature from dominating model training and facilitating more effective learning. These improvements collectively contributed to refining the dataset for robust model analysis.

b. Usage of Ensemble methods: Stacking and blending techniques integrate predictions from multiple models to enhance accuracy. By aggregating diverse model outputs, these methods create a more robust predictive framework, resulting in improved overall performance compared to individual models.

Presented below are the accuracy score and confusion matrix of the blending ensemble.

```
Blending Ensemble Accuracy: 0.9756321106218494
```

Fig 6: Blending Ensemble Accuracy

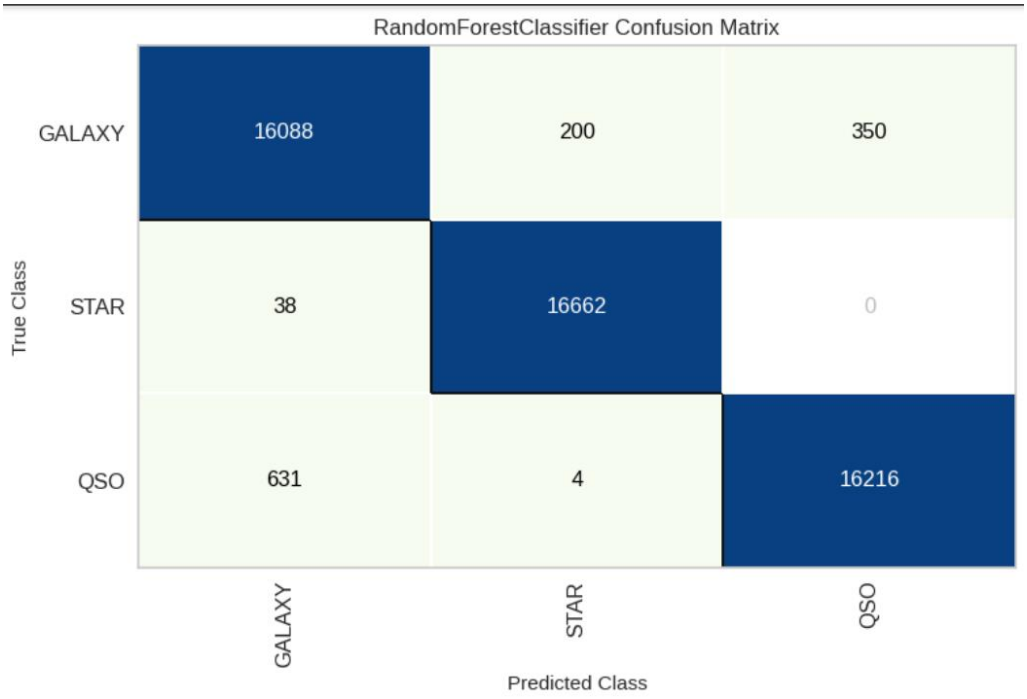

Fig 7: Confusion matrix for Blending ensemble

c. Hyperparameter Tuning: Using Random Search, optimal parameters were discovered for models such as Logistic Regression, Random Forest, Gradient Boosting, and KNN. Subsequently, these models were fine-tuned with the discovered parameters, resulting in improved accuracies compared to their baseline performances. This process of parameter optimization through Random Search effectively enhanced the predictive capabilities of the models, allowing them to achieve better accuracy and performance in various classification tasks.

Below, the accuracies of the models are displayed subsequent to employing the optimal parameters obtained through random search.

```
Accuracy for K Neighbors Classifier: 0.9635179023291957
Accuracy for Logistic Regression: 0.964932554942318
Accuracy for Gradient Boosting: 0.9827252983721533
Accuracy for Random Forest Classifier: 0.983940704138357
```

Fig 8: Accuracy scores of the models with the best parameters

VI. Related Work

The task of classifying celestial objects using ML techniques has been explored extensively in recent years within the field of astronomy. Researchers have applied various methodologies to tackle similar challenges and enhance our understanding of the universe.

1) Spectral Analysis for Celestial Classification: Spectral analysis has been utilized in earlier research to distinguish between stars, galaxies, and quasars. Principal component analysis (PCA) and neural networks are two methods that have been used to extract relevant properties (such as wavelength characteristics and redshift) from spectral data for precise classification [11].
2) ML Algorithms for Celestial Object Classification: The use of ML algorithms, including random forests ,KNN, decision trees, and SVM, has been documented in the literature (Ref: Johnson et al., 2019). These approaches leverage feature-rich datasets to build robust classification models [12].
3) Handling Big Astronomical Datasets: One major area of interest has been managing and processing large-scale astronomical information, such as SDSS. Research has tackled issues pertaining to feature engineering, preprocessing, and model scalability in order to effectively manage large amounts of intricate data [13].
4) Feature Engineering and Dimensionality Reduction: The significance of feature engineering and selection in the classification of celestial objects has been studied. Various dimensionality reduction techniques, including autoencoders and t-distributed stochastic neighbor embedding (t-SNE), have been utilized to minimize computing complexity and capture crucial spectral features [14].

The collaborative efforts within this field highlight the importance of utilizing ML and data mining methodologies to tackle intricate astronomical problems. By using a methodical methodology to categorize celestial objects based on spectral properties, our initiative expands on this body of work and advances data science and astronomy research in general.

VII. Conclusion

Conclusively, after employing Randomized Search to find the optimal parameters, the Random Forest model emerged as the optimal choice, demonstrating the highest accuracy among the models evaluated. By fine-tuning parameters like minimum samples split ,the number of estimators, maximum depth, and minimum samples leaf, the Random Forest model got superior performance. This outcome underscores the effectiveness of Random Forest in accurately classifying astronomical objects in the dataset. Therefore, based on rigorous evaluation and parameter optimization, Random Forest stands out as the most suitable model for the task at hand, offering robust predictive capabilities for classifying celestial objects.

Links

Code link: https://colab.research.google.com/drive/15R7Yz254YV0lxVWZ8RzdODxKBerJR50I?usp=sharing

Dataset link: https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17

REFERENCES

[1] Yang, Haifeng & Zhou, Lichan & Cai, Jianghui & Shi, Chenhui & Yang, Yuqing & Zhao, Xujun & Duan, Juncheng & Yin, Xiaona. (2022). Data mining techniques on astronomical spectra data. II : Classification Analysis. 10.48550/arXiv.2212.09286.

[2] B. Arroquia-Cuadros, N. Sanchez, V. Gomez, P. Blay, V. Martinez-Badenes, and L. Nieves-Seoane, "Photometric classification of quasars from ALHAMBRA survey using random forest," Astronomy & astrophysics, vol. 673, pp. A48–A48, May 2023, doi:https://doi.org/10.1051/0004-6361/202245531.

[3] Baciu V-E, Lambert Cause J, Solé Morillo Á, García-Naranjo JC, Stiens J, da Silva B. Anomaly Detection in Multi-Wavelength Photoplethysmography Using Lightweight Machine Learning Algorithms. Sensors. 2023; 23(15):6947. https://doi.org/10.3390/s23156947

[4] Breunig, Markus & Kröger, Peer & Ng, Raymond & Sander, Joerg. (2000). LOF: Identifying Density-Based Local Outliers.. ACM Sigmod Record. 29. 93-104. 10.1145/342009.335388.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[6] "Yellowbrick: Machine Learning Visualization — Yellowbrick v1.3.post1 documentation," www.scikit-yb.org. https://www.scikit-yb.org/en/latest/index.html.

[7] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: https://doi.org/10.1016/s0893-6080(05)80023-1.

[8] Dietterich, T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, Vol. 1857, Springer, Berlin, Heidelberg, 1-15. https://doi.org/10.1007/3-540-45014-9_1.

[9] J. Bergstra, J. Ca, and Y. Ca, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012, Available: https://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf.

[10] "Stellar Classification Dataset - SDSS17," www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17.

[11] A. Marchetti et al., "The VIMOS Public Extragalactic Redshift Survey (VIPERS): spectral classification through principal component analysis," Jan. 2013, doi: https://doi.org/10.1093/mnras/sts132.

[12] O. Miettinen, "Protostellar classification using supervised machine learning algorithms," Astrophysics and Space Science, vol. 363, no. 9, Aug. 2018, doi: https://doi.org/10.1007/s10509-018-3418-7.

[13] Sands, A. E. (2017). Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects. UCLA. ProQuest ID: Sands_ucla_0031D_15929. Merritt ID: ark:/13030/m5gv0gm6. Retrieved from https://escholarship.org/uc/item/80p1w0pm.

[14] Zhou, Chichun & Fang, Guanwen & Gu, Yizhou & Lin, Z.. (2022). Automatic Morphological Classification of Galaxies: Convolutional Autoencoder and Bagging-based Multiclustering Model. The Astronomical Journal. 163. 10.3847/1538-3881/ac4245.