International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 11, Issue 2 (2025)
https://doi.org/10.31436/ijpcc.v11i2.537

Automated Classification of Celestial Objects Using Machine Learning

Muhammad Aiman Haris bin Muhamad Suwaid[1], Muhammad 'Ilyas Amierrullah bin Ab Karim[2], Raini binti Hassan[3]*, Azni binti Abdul Aziz[4]

[1,2,3]Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia, Kuala Lumpur, Malaysia.
[4]Department of Physics, Kulliyyah of Science, International Islamic University Malaysia, Pahang, Malaysia

*Corresponding author: hrai@iium.edu.my
(Received: 6th January 2025; Accepted: 20th May, 2025; Published on-line: 30th July, 2025)

Abstract— The swift expansion of astronomical data requires the automated classification of celestial objects for practical use. Because of its manual and monotonous nature, classification is more prone to errors and is rapidly losing its viability. This study performs the classification of stars, galaxies, and quasars from SDSS (Sloan Digital Sky Survey) data using the Random Forest, XGBoost, Decision Tree, Gradient Boosting, Linear SVM, KNN, and Logistic Regression. In order to fix the imbalance in the data, the SMOTE algorithm was used, making the model more robust. Random Forest topped the models with their accuracy and reliability across many multiple data releases, hitting an astonishing 99.12% accuracy in SDSS DR18. This work shows how much machine deep learning can change astronomical surveys, providing readily available, precise techniques that are much more effective than manual approaches. The results add towards the development of astrophysics while simultaneously meeting Sustainable Development Goal 9 on fostering innovation through the need for infrastructure.

Keywords— SDSS, Astronomy, Machine Learning, Random Forest, Classification.

I. Introduction

Due to mobility, professional astronomers encounter many challenges because of the exponential growth in astronomical data, especially in the area of classification of objects in the sky. The process of identifying celestial objects using conventional methods, such as human identification and classification methods based on the observable characteristics, is becoming extremely difficult and unfeasible due to the vast amount of data captured by modern telescopes such as the Sloan Digital Sky Survey (SDSS). Through multiple SDSS data release, including DR12, DR14, DR16, DR17 and DR18, had supplied us with extensive spectroscopic and photometric datasets, with over 400,000 celestial objects recorded across these data releases. All of the datasets releases offered detail and rich datasets. This enables researchers to analyze the structure, motion and composition of astronomical entities with greater accuracy and efficiency [23], [26]-[29]. However, these manual classification approaches are not only labour intensive but are also prone to human error. This will make it harder for professional astronomers to effectively gain valuable insights from the available data for more scientific discoveries and advancement [1][9].

Today, the amount of data is growing rapidly, which causes the traditional classification method used by astronomers to become insufficient due to the slow process of classification and high probability of making errors [1][9]. There is a requirement for designing and implementing the automated classification systems to fulfil the requirements of astronomers in analysis and classification of very large data accurately [7][8]. This matter should be addressed as soon as possible because it slows down the developments in astronomical science and creation of new inventions [3],[11].

The aim of this work is to design a reliable machine learning model capable of classifying objects such as galaxies, QSOs, or stars. The goal of the research is to train classifier capable of handling the rich nature of astronomical information by using large dataset of SDSS from several data release. This type of model is important to construct because of the drawbacks of conventional classification methods and it will make it easier for astronomers to manage and evaluate a large amount of data.

The research also focuses on assessing and comparing different classifiers that are used to design the machine learning models. Each classifier must be evaluated by the metrics like precision, accuracy, f1-score and recall because to unearth out the most efficient classifier for classifying the astronomical objects.

In the assessment of SDSS, the study deploys the data source to build and test a variety of machine learning models for classifying a variety of astronomical objects. The stages include data preparation, model developing and model assessment that ensure the proper application of classification methods. The algorithms are presented and compared in a detailed yet concise manner, taking into

https://doi.org/10.31436/ijpcc.v11i2.537

account a variety of evaluation metrics, including precision, recall, accuracy, and F1 score.

The main audience of this project is astronomers and academic researchers because they are always active in researching and analysing objects in space using the data they obtain. This study will also benefit institutions related to astronomical research because they always take data and pictures of objects in space using their modern telescopes.

Python language will be used because of its strong support for data science and machine learning where the various libraries available are Scikit-learn, Pandas, Numpy, Matplotlib among others. The computationally demanding experiments will be conducted on good performance computing that is capable of handling large datasets.

This research can enhance the progress of astronomical research because it is able to provide a more effective technique in classifying celestial objects as well as helping astronomers to improve their knowledge and understanding of nature more deeply. This initiative to create a classification model will help the advancement of this field of astronomy.

Furthermore, this initiative will significantly enhance research efficiency by facilitating quick analysis and processing of large amounts of astronomical data. Conventional classification techniques are inadequate due to the rapid increase in data collected by present-day telescopes. The use of machine learning in classifying celestial objects can speed up the classification process as well as save time and resources and give astronomers the opportunity to focus on more complex studies that require more effort. This will increase their productivity and facilitate their process of making new discoveries in the astronomy field.

II. LITERATURE REVIEW

The proposed study builds insights and methods from the reviewed literature to create an effective framework for classifying celestial objects into stars, quasars, and galaxies (see Appendix 1 and 2). Key adaptations and improvements have been made at various steps of the process to enhance performance and address limitations identified in prior studies. They make the study both reliable and efficient in terms of the methodology to be used in the study.

This dataset was sourced from the Sloan Digital Sky Survey as used by a number of authors such as Solorio-Ramirez et al. [5] and Er and Bilgin [33]. The selection is further supported by Zeraatgari et al. [35], who, together with the ALLWISE catalog data, managed to combine it successfully for classification purposes. Furthermore, Cruz et al. [32] reported SDSS is a good reference for spectral classification and accuracy of over 94 percent were carried out using Random Forest and Neural Networks.

Preprocessing remained a key focus as the data had to be prepared to give optimal results before analysis. This was done in order to manage missing data, remove duplicates, and ensure equitable distribution of cases and controls. Following the Hassina [7] and Zhang [34] methodologies, class imbalance was addressed by the application of SMOTE as the primary step of the class balancing which is essential for accurate prediction on the datasets with severely imbalanced categories. This point has been captured further by Er and Bilgin [33] who demonstrated that class balancing improved the accuracy of the predictive model from 87.71% to 94.67%, which is a substantial increase.

Increased efficiency in computation and improvement in model parameters was achieved through selective feature choices. Such conclusions are backed by Sharma and Sharma [9] as well as Vavilova et al. [1] who discussed the significance of making work with the model more effective through feature reduction and irrelevant features removal. Zhang [34] added to this information showing that the narrowing of focus during feature selection particularly to redshift and photometric features boosted the accuracy to 99.39% with XGBoost.

Because all features were expected to contribute to learning, data was normalized following Yoshino et al. [8] and Zeraatgari et al. [35]. Such normalization avoided algorithms, such as SVM, from struggling because of sensitivity to input scale. Exploratory analysis was done to understand the data, align with patterns, trends, or outliers like Smita et al. [11] did.

As quoted by Ashai et al. [2], the dataset was SMOTE oversampled to balance it out. This not only enhanced the model's performance but also provided artificial data for the minority categories which has been approved by Er and Bilgin [33], who did a systematic review on supervised machine learning techniques. These methods allowed for precision modeling with fairly evenly distributed data sets, especially for the more seldom detected quasars and other minority classes.

A set of algorithms were implemented for machine learning to facilitate building a complex classification system. The Random Forest algorithm was included because it is known to be effective with noisy and mixed data [5], [32]. Gradient boosting and XGBoost were selected because of their multi-stage error reducing method [3], [34]. SVM was applied to classify data with large dimensionality [9], [1]. K-Nearest Neighbors (KNN) was chosen because it is effective in classification problems of this sort [2]. Another set of algorithms for classifying cases is based on Decision Trees because they are easy to use, and easy to understand [10]. At last, the Logistic Regression model was employed because, traditionally, it is very effective in classification problems whether binary or multi-class, and it is able to

https://doi.org/10.31436/ijpcc.v11i2.537

provide reliable midpoints in the cases of the galaxy morphological types and could appropriate high levels of accuracy [1].

Although the introduction of Neural Networks (NNs) have successfully accomplished complex tasks, but for the scope of the research conducted in this case, NNs were not the recommended choice. Primarily, NNs will not perform to their best capabilities without ample amounts of data, which greatly exceeds the availability of the SDSS dataset. Secondly, even without considering NNs, models such as Random Forest and XGBoost already claimed remarkable results on previously conducted astronomical classification studies, thus eliminating the requirement of more advanced models [5], [8]. Furthermore, the vast majority of NNs are not as easily interpretable compared to regression trees which was a major concern for this research when it came to analysis of feature importance.

As for focusing on ensemble methods, Random Forest, XGBoost, and Gradient Boosting were selected because they have addressed multilevel class imbalances effectively as the literature review showed (See Appendix 1). AdaBoost and Extra Trees, on the other hand, were not selected because those models will simply serve to add redundancy, and are based on the same ideas as the selected models without outperforming them in other relevant researches [3], [7]. The requirements did not set any constraints and instead concentrated on the relative aspects of interpretability, efficiency of computation, and precision of classification provided by single robust algorithms, which can easily be accomplished through the existing benchmarks in the field.

III. METHODOLOGY

The dataset comes from the images that have been previously taken by the SDSS cameras, and then the pre-processed data was shared on Kaggle. This study focuses on classifying the data into three categories: galaxies, QSOs, and stars.

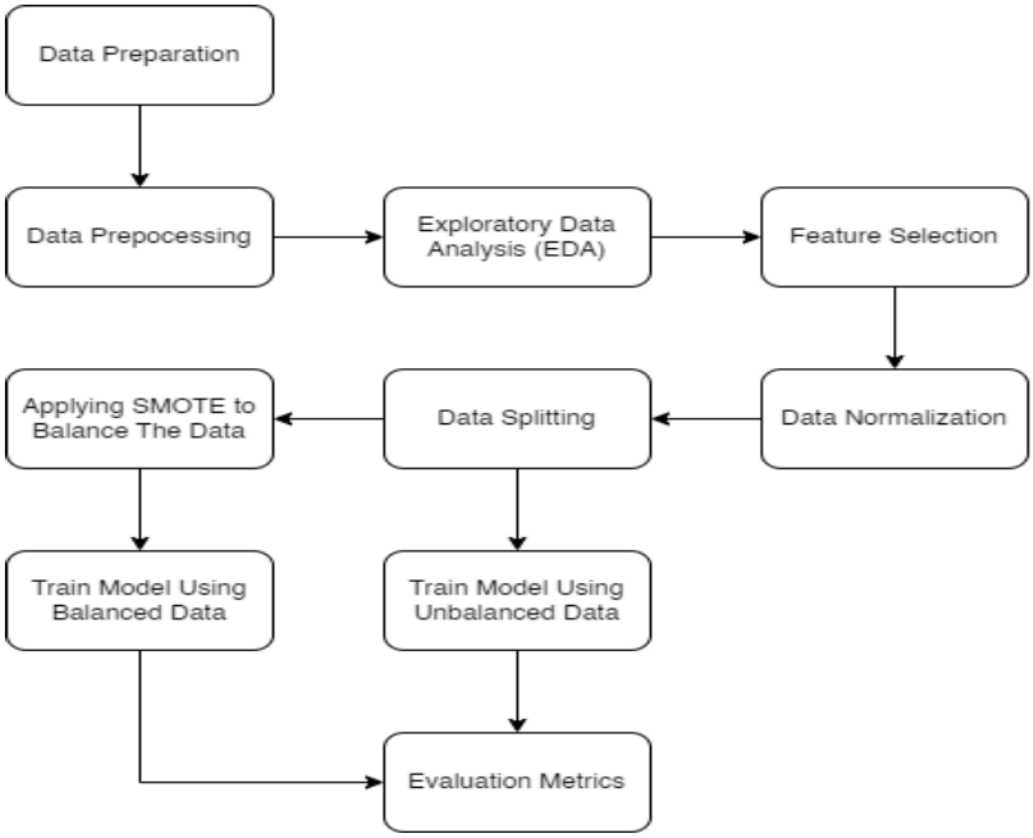

Fig. 1 The process flow of methodology

A. *Tools: Python and Pycharm*

The tools applied to this project were selected in order to implement the intended system for formation of celestial objects classification using machine learning. Python was the primary coding language due to variety and availability of numerous valuable libraries for data science and machine learning streams. Dealing with the dataset was facilitated by Pandas while NumPy was used for operations such as operation on the arrays and matrices. A package, Matplotlib,

https://doi.org/10.31436/ijpcc.v11i2.537

was used for the plotting and visualizing of data, as a way of making sense of patterns into the data generated.

B. *Dataset*

This dataset for this project was obtained from the Sloan Digital Sky Survey (SDSS) Data Releases 12, 14, 16, 17, and 18 and includes 418,070 rows of spectroscopic and photometric data concerning celestial objects. SDSS is a large-scale survey conducted at the Apache Point Observatory in New Mexico, using a specialized 2.5-meter-wide angle optical telescope. This survey has collected data on millions of celestial objects, providing deep, multicolour photographs that cover one-third of the sky [24].

This project focused on three types of celestial objects: galaxies, stars, and quasars. These were chosen because they emit their own light, making them easier to observe compared to objects like planets, gases, or black holes that rely on external light sources to be visible. By narrowing the scope to these three categories, the dataset became more manageable and relevant for building and evaluating machine learning models. The richness and detail of the SDSS dataset made it an ideal choice for this research, providing a strong foundation for accurate and meaningful classification [1].

C. *Data Preprocessing*

Data preprocessing was started with the first step, which is loading the dataset into a Pandas DataFrame. First, it was done to gain insight into the structure of the obtained data set. The researcher adopted the use of the command df.head() to get a preview of the content in a DataFrame. Thus, df.dtypes was used to check if each of the columns contained integer, floats, string or object values. In order to learn about the structure of the dataset containing integers, floats, strings or objects, the command df.shape was used.

The next challenge was how to handle missing values, which are detrimental to many machine learning models. These were detected using df.isnull().sum() command, on the data frame and then they were excluded by using the statement df.dropna(inplace=True). Rows were repeated which could bring a bias and redundancy, this was sorted out using df.duplicated() and removed using df.drop_duplicates(inplace=True). The target variable was basically in the form of categorical attributes where distributions and objects were classified as 'GALAXY,' 'STAR,' and 'QSO.' For machine learning purposes, numerical labels were given to each of these categories. In particular, 'GALAXY' was equated with '0,' 'STAR' with '1' and 'QSO' with '2.' Preprocessing steps have made the clean and effective preparation of dataset suitable for training the machine learning algorithm.

D. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) was carried out to better understand the dataset and visualize the relationships between features. The main steps included:

1. Feature Distributions:
- Histograms were created to show how the values of each feature are spread across the dataset for different data releases.
- Boxplots were used to detect outliers and understand the range of each feature, helping to determine if outliers were meaningful or caused by errors.

2. Pairwise Relationships:
- Pair plots were generated to show the relationships between features. These plots helped us see how features like photometric bands and redshift are related and whether there are clusters or patterns in the data.

3. Correlation Matrix:
- Numerical features' association strength checking was based on the use of a correlation matrix. This assisted in detecting features whose relation is high and 'broadly' similar to the correlations identified between certain photometric bands (like r and i) and redshift adds additional information.

Such steps allowed receiving useful information about the given dataset and selecting features which are more significant for classification of celestial objects. This analysis was useful in the process of building this model and enhancing the outcomes of the research as well.

E. *Feature Selection*

Feature selection is a critical step in machine learning process and plays an enormous role in the improvement of input signals. The following is because, it saves time for computation, it is less likely to over-fit, and it is suitable for other unseen data. To reduce the dimensionality of the data set a process of feature selection was conducted to remove unneeded columns and features with variance of zero and the final dataset only contained features that were meaningful for classification of celestial objects.

The first operation performed was to detect features with feature variance equal to zero and then eliminate them. Variance can be defined as statistical measure of dispersion that quantifies how far apart from the mean of a feature values are. Features that have variance of zero have the same value for all observation which makes them unhelpful when doing predictive modelling. The features were flagged as comprising two variables: objid and rerun; these were excluded from the data set.

Other columns that were dropped since they contributed little to the classification of astronomical objects include

https://doi.org/10.31436/ijpcc.v11i2.537

features with zero variance. Often, these columns included miscellaneous information or ID's, etc., that excluded from the observation layer; in other words, the column could be less relevant during the classification process. The following features were excluded from the dataset:

- specobjid: Spectroscopic object identifier
- ra: Right ascension of the object
- dec: Declination of the object
- run: SDSS imaging run identifier
- camcol: Camera column
- field: Field number
- plate: Plate number
- mjd: Modified Julian Date
- fiberid: Fiber ID

Id-depending information and observational meta-information and extra information on the photometric or morphological aspect were taken out of the dataset to classify the feature space as simple as possible. The last subset of features chosen for the model is redshift together with u, g, r, i, z photometric bands. This was done to counter the observed variability in feature selection found in the current literature. However, these specific features such as redshift and photometric bands were applied persistently in many studies [3], [5], [7], [8]. While other features were selected, the justification for doing so was not always comprehensible, and the same can apply to the repeated use of this strategy. These theses prove the significance of these features for the classification of celestial objects. Redshift gives important data about distance and velocities of objects and with photometric bands that give information about the spectrum energy distribution which help to distinguish between stars, galaxies and quasars.

F. *Data Normalization*

In this study, all numerical features, other than the target variable class, were normalized using the StandardScaler function from the sklearn.preprocessing library. This method made each feature to have mean equal to zero and the standard deviation equal to one [25]. The cases include algorithms like Support Vector Machines (SVM) where standardization must be performed due to its scale of input data. It also stops feature with large value dominating features with small value, thus all numerical features are given an equal input into making the model.

G. Data Splitting

The data was split into two parts: X and y. The variable X was a combination of all features after normalization except the class and y was the target variable class which contained the labels for celestial objects. Using the function in sklearn. model_selection which is train_test_split, the data was split into train data at 70% and test data at 30%. Another configuration for the state of each component was a purely random choice of 42 in order to achieve a reproducible result. This split created four parts: These are; X_train (training features), y_train (training target), X_test (testing features), and y_test (testing target).

The 70/30 split is widely used in machine learning as it is effective in balancing between the need for a sufficient training data to optimize model and to adequate testing data to evaluate its performance. Empirical studies shown that by allocating 20 – 30% of data for testing provides the best trade off between minimizing approximation errors and also maximizing the validity of the model results. So, this ensures that the model generalizes well to unseen data while avoiding overfitting, as highlighted in [30].

K-fold cross validation provides a more robust approach as it uses the entire dataset for training and validating in multiple iterations. However, it can be computationally expensive especially for the large datasets like the one used for this study. So, the 70/30 split is computationally efficient and suitable for this project after considering the size of SDSS dataset and the available computational resources.

H. *Applying Synthetic Minority Over-Sampling Technique (SMOTE)*

Using Synthetic Minority Over-Sampling Technique (SMOTE) to handle imbalanced dataset, where one class has the most number or frequency compared to other class. In this study, SMOTE was applied using the imblearn library. The fit_resample method was used on the training data (X_train and y_train) to generate new data for the minority class, making the dataset balanced.

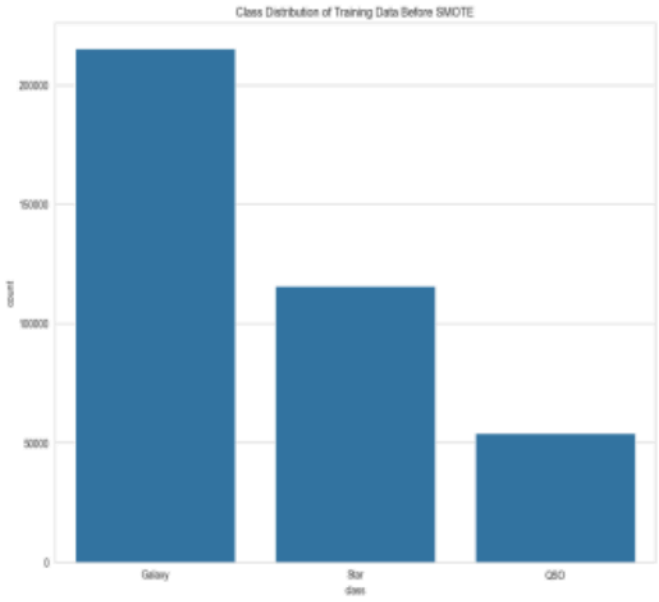

Fig. 2 Class distribution of training data before SMOTE

https://doi.org/10.31436/ijpcc.v11i2.537

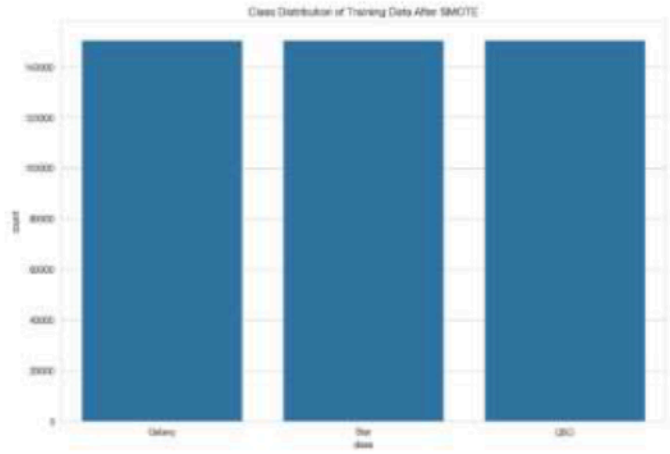

Fig. 3 Class distribution of training data after SMOTE

I. *Modelling*

The model was trained using seven different machine learning algorithms on raw and balanced data to assess and understand their behaviour under different data conditions, each with its own approach to classification:

- Gradient Boosting builds models iteratively, with each new model minimizing the errors of the previous one, combining them into a strong learner [16].
- XGBoost is an optimized version of Gradient Boosting designed for speed and efficiency, using ensemble learning to improve predictions [14][15].
- Decision Tree uses a tree-like structure where nodes represent decisions, branches represent outcomes, and leaves represent predictions, splitting the data based on selected features [13].
- Linear SVM finds the optimal decision boundary (hyperplane) that separates data points into different classes in feature space [19].
- Random Forest creates multiple decision trees using subsets of data and combines their outputs for a final prediction [18].
- K-Nearest Neighbours (KNN) predicts the label of a data point by analysing the labels of its closest K neighbours based on distance [17].
- Logistic Regression models the relationship between variables to calculate the probability of a data point belonging to a specific class [12].

Although the introduction of Neural Networks (NNs) have successfully accomplished complex tasks, but for the scope of the research conducted in this case, NNs were not the recommended choice. Primarily, NNs will not perform to their best capabilities without ample amounts of data, which greatly exceeds the availability of the SDSS dataset. Secondly, even without considering NNs, models such as Random Forest and XGBoost already claimed remarkable results on previously conducted astronomical classification studies, thus eliminating the requirement of more advanced models [5][8]. Furthermore, the vast majority of NNs are not as easily interpretable compared to regression trees which was a major concern for this research when it came to analysis of feature importance.

As for focusing on ensemble methods, Random Forest, XGBoost, and Gradient Boosting were selected because they have addressed multilevel class imbalances effectively as the literature review showed (Appendix I). AdaBoost and Extra Trees, on the other hand, were not selected because those models will simply serve to add redundancy, and are based on the same ideas as the selected models without outperforming them in other relevant researches [3][7]. The requirements did not set any constraints and instead concentrated on the relative aspects of interpretability, efficiency of computation, and precision of classification provided by single robust algorithms, which can easily be accomplished through the existing benchmarks in the field.

J. *Model Evaluation*

The evaluation of the machine learning models was performed using several key metrics to ensure a comprehensive understanding of their performance:

1) Accuracy: This metric represents the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of instances in the dataset. Accuracy is particularly useful for providing a general understanding of model performance but can be misleading in the presence of class imbalance.
2) Precision: Precision measures the ratio of true positive predictions to the total number of predicted positives. This metric is particularly important in cases where minimizing false positives is crucial, such as identifying quasars from other celestial objects.
3) Recall: Recall, also known as sensitivity, calculates the ratio of true positive predictions to the total actual positives. This metric is crucial for understanding how well the model identifies all instances of a particular class, especially for rare classes like quasars in the dataset.

https://doi.org/10.31436/ijpcc.v11i2.537

4) F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure that is particularly useful when there is a significant class imbalance in the dataset. This metric ensures that both precision and recall are considered equally in the evaluation.

5) Confusion Matrix: A confusion matrix was generated to provide a detailed view of the model's performance across each class. It shows the counts of true positives, false positives, true negatives, and false negatives for each class (galaxies, stars, and QSOs). This matrix is essential for understanding the specific misclassification patterns and identifying areas where the model could be improved.

The evaluation was conducted on both the raw and balanced datasets, ensuring that the impact of class balancing techniques, such as SMOTE, was considered in the analysis. The metrics were calculated for each of the three classes—galaxies, stars, and quasars—individually to evaluate the model's performance across different celestial object categories comprehensively. This multi-metric evaluation approach ensured a robust comparison of the models and helped in identifying the best-performing algorithms for the classification of celestial objects using the SDSS data.

K. *Statistical Analysis*

Stastical analysis were conducted in order to validate the performance comparisons between each machine learning models. The analysis was done according to a detailed statistical framework. This framework combined parametric and non-parametric statistical test to validate the comparative performance of the models acrross multiple datasets of data release. The analysis was guided by the framework outlined by Chatzi and Doody [31] which emphasizes testing assumptions before choosing the appropriate statistical methods.

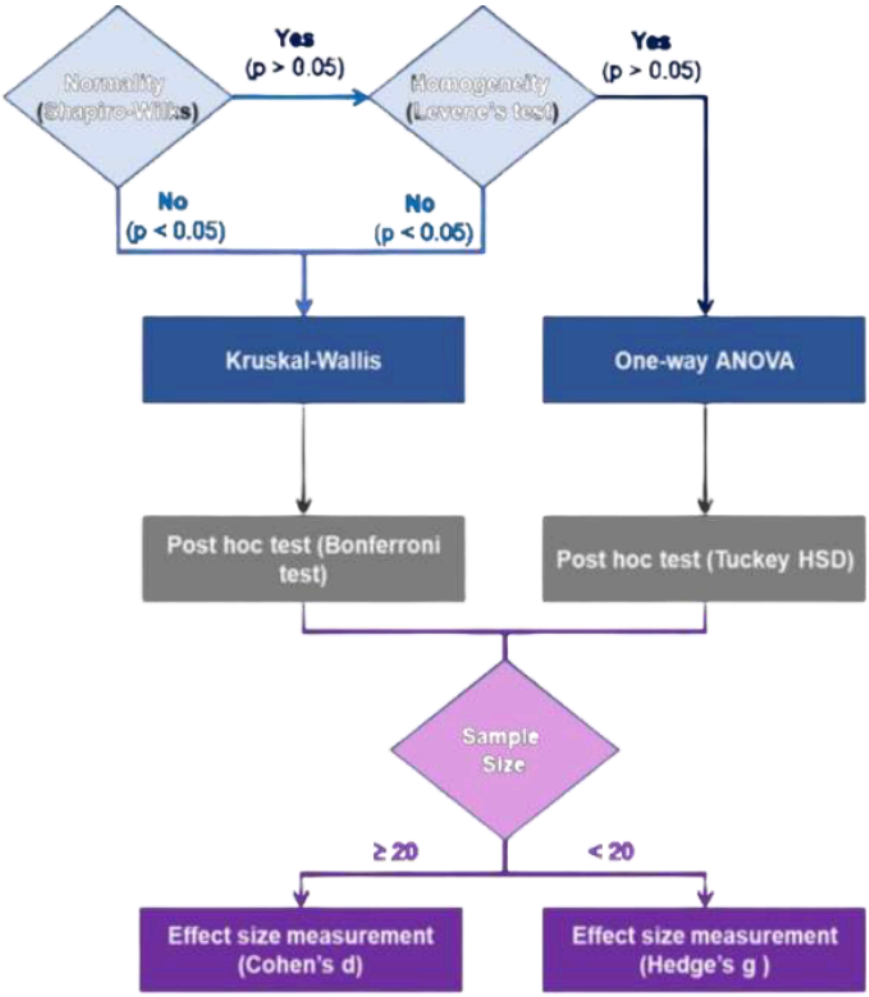

Fig. 4 Framework for selecting statistical analysis methods. Adapted from [31].

https://doi.org/10.31436/ijpcc.v11i2.537

1) Normality Assessment

The distributions of accuracy for each machine learning models were examined using the Shapiro-Wilk's test. A threshold of 0.05 was used to determine whether the data was significantly distributed and thus normal. Models that had a p value less than or equal to 0.05 were considered as having non-normal distributions. The findings clearly demonstrated that all models did not exhibit normality and some of them particularly Gradient Boosting and KNN did not satisfy the normality assumption. As discussed [31], this normality testing was important in deciding whether to use a parametric or a non-parametric test.

2) Homogeneity of Variance

Levene's test was performed to assess the equality of variances among the models' accuracies. The test results ($p \geq 0.05$) confirmed that homogeneity of variance. This fulfilled one of the criteria to allow the application of parametric. Variance homogeneity is one of the key assumption for parametric tests such as Analysis of Variance (ANOVA) and was addressed to ensure accurate comparison, consistent with [31]. However, if the normality assumption fails then, then analysis can only proceed to non-parametric tests.

3) Group Comparisons

- This test was done to check the differences of mean accuracy value across models when both normality and homogeneity assumptions are satisfied. According to [31], ANOVA is perfectly able to compare more than two means at one time and eliminates Type I error.
- Kruskal-Wallis Test:
  For those models which did not meet normality assumptions, the Kruskal-Wallis non-parametric test was used to detect any deviation from median accuracy. The p-value for this test was set at 0.05. Such results indicate that at least one model did not agree with the rest. This concurs with [31] warnings on the use of non-parametric methods where assumptions are known to fail.

4) Post-Hoc Analysis

Post hoc analysis aims to pinpoint the model which has impacted the research in question the most. [31] suggested that for Kruskal-Wallace test, Dunn's test with Bonferroni correction should be made for pairwise difference decision to eliminate Type 1 error. For ANOVA, the difference of means HSD test will assist in figuring out pairwise differences between models.

IV. RESULTS

A. *Comparative Performance Across SDSS Data Releases*

Figure 5 depicts the comparative performance of various machine learning algorithms—Random Forest, XGBoost, Gradient Boosting, Decision Tree, Linear SVM, KNN, and Logistic Regression—across multiple Sloan Digital Sky Survey (SDSS) data releases (DR12 to DR18). The results presented offer a comprehensive analysis of the models' accuracies under varying dataset complexities, shedding light on their adaptability and robustness.

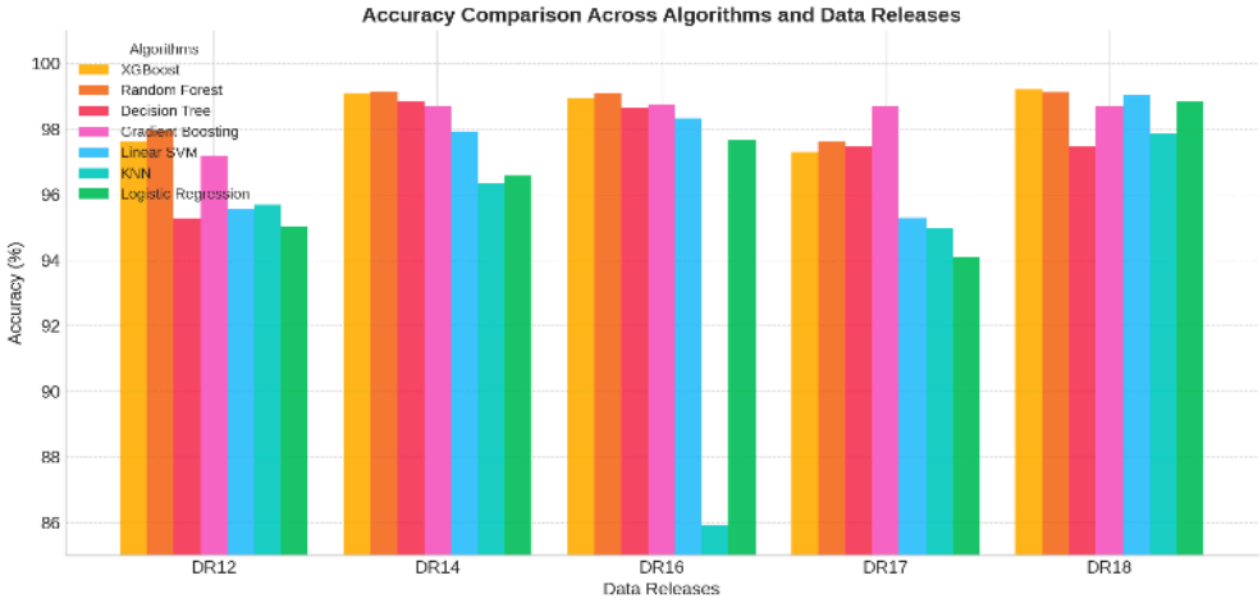

Fig. 5 Comparison of Accuracy Across Different Machine Learning Algorithms (e.g., Random Forest, XGBoost) for Classifying Celestial Objects Across SDSS Data Releases (DR12 to DR18)

https://doi.org/10.31436/ijpcc.v11i2.537

Random Forest consistently excels across all data releases and emerges as the most dependable algorithm. Its peak performance of 99.12% on DR18 validates its robustness and highlights its scalability to handle complex and diverse datasets. Its reliability in earlier datasets like DR12 (97.95%) further underscores its consistency. These findings corroborate previous research, such as [5], emphasizing the algorithm's capability to manage noisy, imbalanced data environments.

XGBoost, closely trailing Random Forest, demonstrates exceptional capability in minimizing misclassifications and managing high-dimensional datasets. Its peak accuracy in DR18 with 99.21% accuracy and competitive performance across earlier releases reinforce its versatility. The algorithm's strong adaptability to different celestial object classes suggests its potential as an alternative when precision is crucial, especially in classifying stars and galaxies.

Gradient Boosting, while achieving an accuracy of 98.69% in DR18, illustrates a trade-off between accuracy and interpretability. Its limitations become apparent in datasets with overlapping spectral features, such as DR16, where minor declines in accuracy were noted. Despite these drawbacks, its balanced performance across multiple datasets makes it a viable option for tasks requiring a compromise between transparency and performance.

Although decision Tree models are efficient for straightforward tasks like star classification, they face challenges in handling complex boundaries, such as those required for quasars. Their reduced F1 scores and accuracy dips in DR17 and DR18 highlight these limitations. Nevertheless, Decision Trees can serve as foundational elements in ensemble methods, offering simplicity and interpretability.

Linear SVM, with accuracies consistently exceeding 94%, struggles to address non-linear separability. This limitation, particularly evident in the classification of quasars, calls for kernel-based enhancements or feature transformations to boost its effectiveness.

Logistic Regression, maintaining a baseline accuracy above 94% across all data releases, highlights its simplicity and reliability for less complex tasks. However, its inability to handle intricate datasets underscores the need for more sophisticated models in applications involving substantial feature overlap.

K-Nearest Neighbors (KNN) recorded the lowest performance among the evaluated algorithms, which underscores its sensitivity to high-dimensional spaces and feature overlap. This trend is most apparent in DR16, where noticeable accuracy drops were observed. These findings emphasize KNN's limitations for large-scale and complex datasets, such as those in SDSS.

The results reveal practical implications for astronomical classification tasks:

1. Ensemble Methods: Random Forest and XGBoost are ideal candidates for automated classification pipelines in large astronomical surveys due to their scalability and robustness.
2. Simpler Models for Benchmarking: Logistic Regression and Decision Trees, while less suitable for complex classifications, serve as valuable benchmarks for evaluating advanced algorithms.
3. Dataset Characteristics Matter: The consistently superior accuracy in DR18 suggests that improved data quality and volume significantly enhance model performance. This underscores the importance of well-curated datasets in achieving optimal results.

Figure 15 confirms that ensemble methods like Random Forest and XGBoost dominate in performance, meeting the study's objectives by providing reliable and efficient solutions for celestial object classification. Meanwhile, the limitations of simpler models like KNN and Logistic Regression reinforce the necessity of advanced techniques for managing the complexities of astronomical data. These findings not only validate the research's approach but also provide a clear roadmap for future explorations in automated astronomical data analysis.

B. *Robustness and Variability of Model Accuracy*

Figure 6 provides a complete breakdown of how each machine learning algorithms' accuracy differed with respect to the SDSS data versions under consideration. This particular analysis demonstrates parallels and inconsistencies of algorithm efficiency, thus revealing how each model reacts to the varying datasets.

Random Forest: The IQR of Random Forest is narrow suggesting that it is consistent with achieving high accuracy score in all of the data versions. In addition, its lack of variability means that it is quite resilient to changes in data. It is worth mentioning that the usage of random outliers is almost non-existent. This attribute of random forest enhances its generalization capabilities on different sets of astronomical data.

K-Nearest Neighbors: Judging by KNN, the narrowest IQR implies that all other algorithms exhibit much broader outliers. This means that KNN is highly dependent on the feature overlaps or class imbalances in the dataset. The presence of multiple outliers in the controlled experimentation means the algorithm is performing inconsistently at bestangement pointing towards the proximity based classifiers for larger matrices that are highly dimensional like SDSS ones has their disadvantages.

https://doi.org/10.31436/ijpcc.v11i2.537

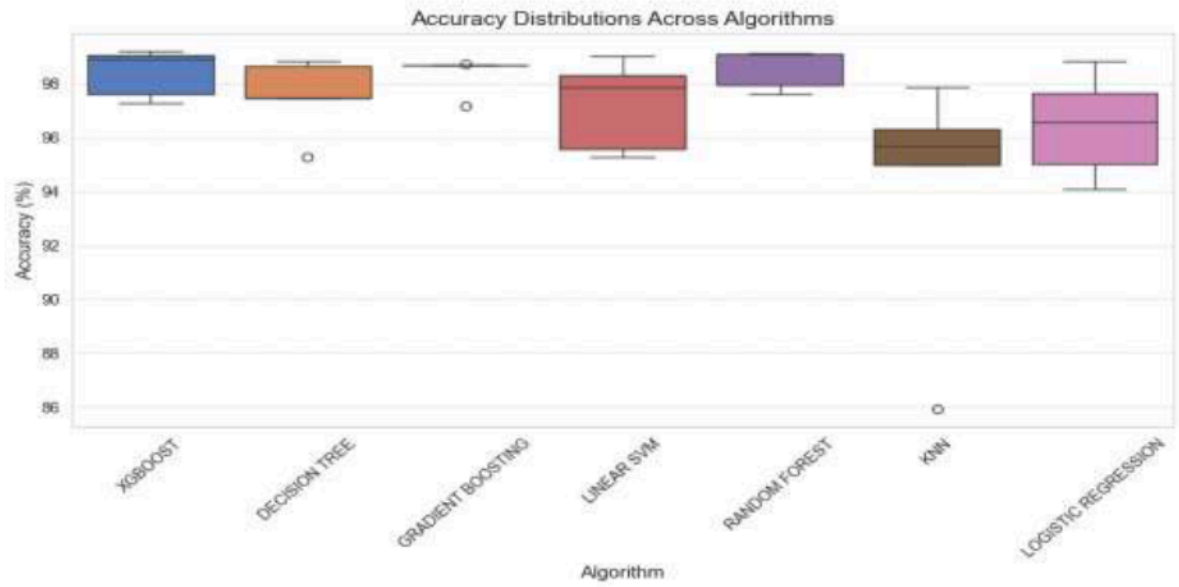

Fig. 6 Accuracy Distributions Across Machine Learning Algorithms

Logistic Regression: KNN is captured by the least variability against the other algorithms. Logistic Regression maintains a moderate median accuracy score while exhibiting central measures variability. This assumption stems from the model's linear structure.

C. *Statistical Validation of Performance Differences*

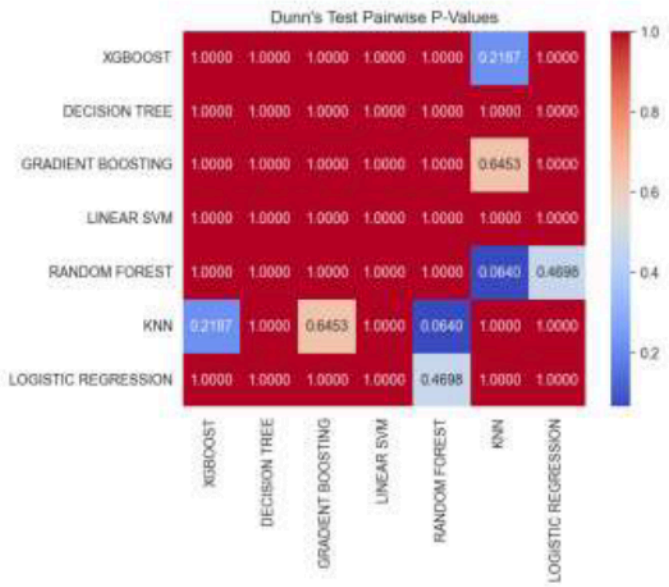

Fig. 7 Dunn's Test Pairwise P-Values Heatmap

Dunn's test results in Figure 7 elucidate the alignment and differences in the accuracy outcomes of different machine learning algorithms. As can be seen from the Random Forest's results, it's performance is quite different when analyzed through the use of K-Nearest Neighbors (KNN) thus showcasing its effectiveness and ability to generalize more robustly. This correlates with the results aim of finding a model that can classify celestial objects accurately and with great certainty.

Practically speaking, this study illustrates that Random Forest is optimal with respect to the processing of large volumes of astronomical data where the dimensionality and sparsity of the dataset is particularly high. The lack of significant differences between the XGBoost and the other algorithms results indicates that Decision Tree and Gradient Boosting may also be adopted as other reasonable methods depending on the nature of the data to be analyzed or the available computational power.

These results highlight the utilization of advanced statistical techniques such as Dunn's post-hoc test in measuring differences in the performance of various algorithms. As stated earlier, this helps ensure the conclusions reached have both statistical and practical value and sets the stage for further research into ensemble techniques and hybrids that would benefit from different algorithms. In so doing, this particular study provides powerful closure to the insights gained by melding these statistical results with the core objectives of the study. This, in turn, culminates to enhancing the techniques employed in automated classification of celestial objects.

The statistical analysis was done following a strict multi-level approach in order to ensure reliability and validity of results with the claims made:

https://doi.org/10.31436/ijpcc.v11i2.537

1) Normality Testing (Figure 18):

As shown in figure 8, the Normality of the accuracy distributions for each machine learning technique was carried out with Shapiro-Wilk test. This was necessary for deciding the use of parametric or non-parametric statistical techniques in the next phases of analysis. From the results, the assumption of normality was not uniformly met by the algorithms.

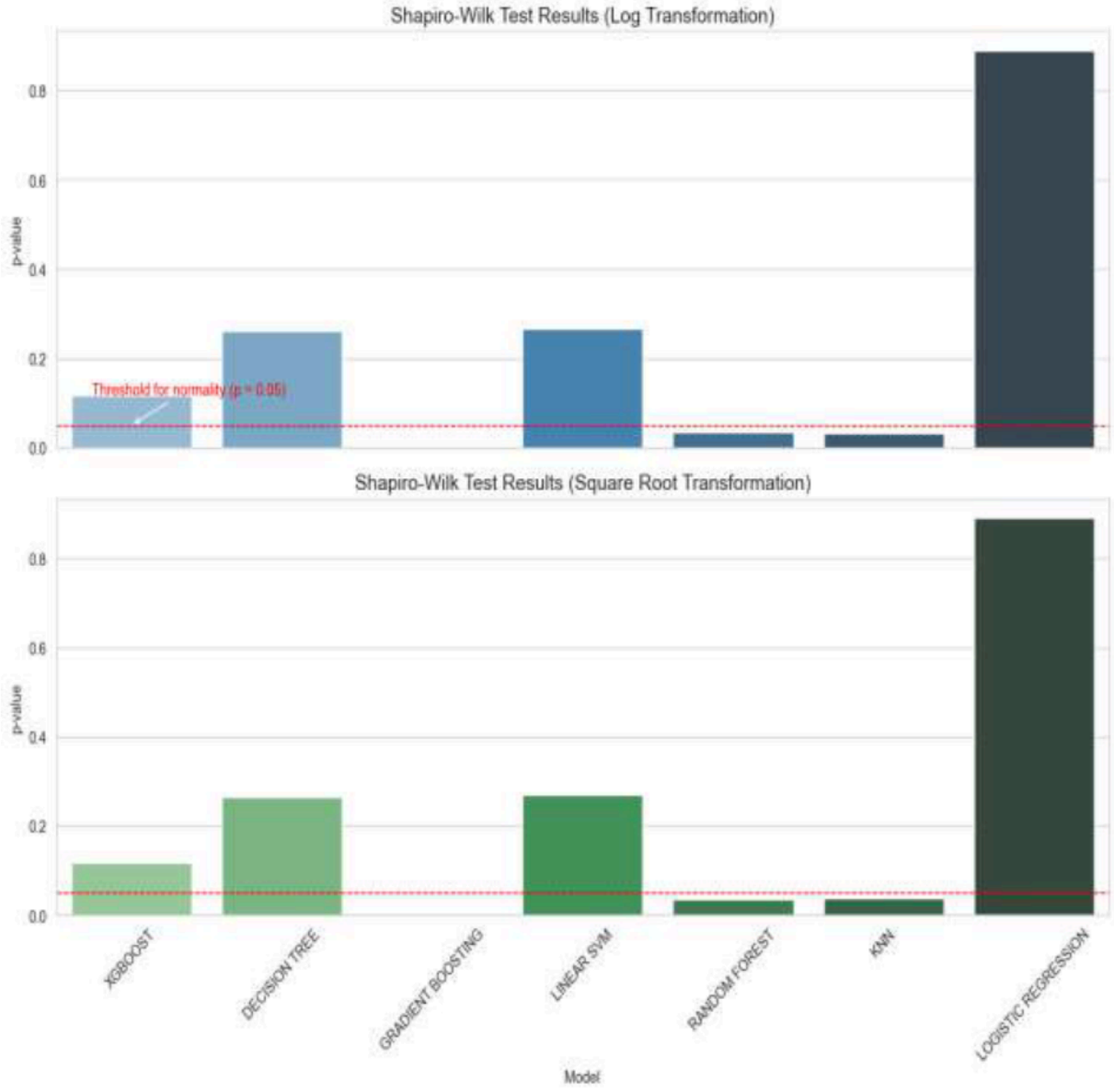

Fig. 8 Shapiro-Wilk Test Results for Evaluating Normality After Log and Square Root Transformations

In particular, it was noted that certain algorithms such as Gradient Boosting, Random Forest, and KNN had a p-value lower than 0.05, indicating a significant result as well as a non-normal distribution. These patterns denote that the precision metrics of these models are subject to dataset composition and model and thus do not permit parametric testing without some form of preconditioning on the data.

On the other hand, algorithms like XGBoost, Decision Tree, Linear SVM, and Logistic Regression had a p-value greater than 0.05, supporting the notion that the accuracy distributions of these models are normal. These findings represent how the algorithms react differently to the dataset and emphasize the need to apply appropriate statistical models according to the specific data set attributes. Log-transformed and square root transformed data for normality testing results have been summarized in Figure 18. The figure highlights the differences in the p-

https://doi.org/10.31436/ijpcc.v11i2.537

values across algorithms and makes it easier to gauge the normality testing results.

2) Variance Homogeneity:

Levene's test was conducted to check if the results of the algorithms tested had a common variance within them, and from the test, it generated a p-value of 0.4241. This implies that the percentage variation between the accuracy scores in the models is statistically homogeneous which allows proceeding with additional analyses.

3) Kruskal-Wallis Test:

As a result of non-mean distribution detected from the Shapiro-Wilk test, the non-parametric alternative for one-way ANOVA was employed and proved effective. Accuracy when performing the algorithms was different among algorithms differing significantly, H-statistic 13.3028 and p-value 0.0385. This data proves that there is sufficient variation in the performance of the model to justify further examination through pairwise analysis.

4) Dunn's Post-Hoc Test:

In order to determine which algorithm pairs have a significant difference, Dunns test was examined with the heatmap generated by the Bonferroni correction, and its results showed difference between Random Forest and KNN (p=0.640). These results further support the notion that Random Forest is indeed more accurate than KNN and more importantly, that the difference is statistically significant. The increase in accuracy demonstrates that Random Forest is able to perform well with a broader range of datasets.

These significant differences in the performance of Random Forest and the other models clearly demonstrate that the algorithm does a commendable job at classifying celestial objects. It is vital to note that, despite its success, models such as KNN are proficient in classification, albeit with much lower accuracy. The absence or rather lack of significant differences between the rest of the models above mentioned proves that maybe those models have distinct values for comparison but would also greatly depend upon the scope of the projects and budgetary allocations for computational resources.

Such detailed scrutiny about Random Forest simply strengthens assumption that it is the best out did not consider one important aspect, which is, differencing out the outlier portion of these other models and making sense of the variance through analysis of variance is beneficial towards understanding machine learning on astronomy.

D. *Class-Specific Classification Metrics*

With respect to the classification of galaxies, stars, and QSOs in SDSS DR18, Figure 9 illustrates the performance comparison of Random Forest and XGBoost with respect to precision, recall and f1-scores. Such metrics provide further insight into the respective capabilities of the techniques with regard to the management of astronomical data, particularly in the presence of imbalanced classes

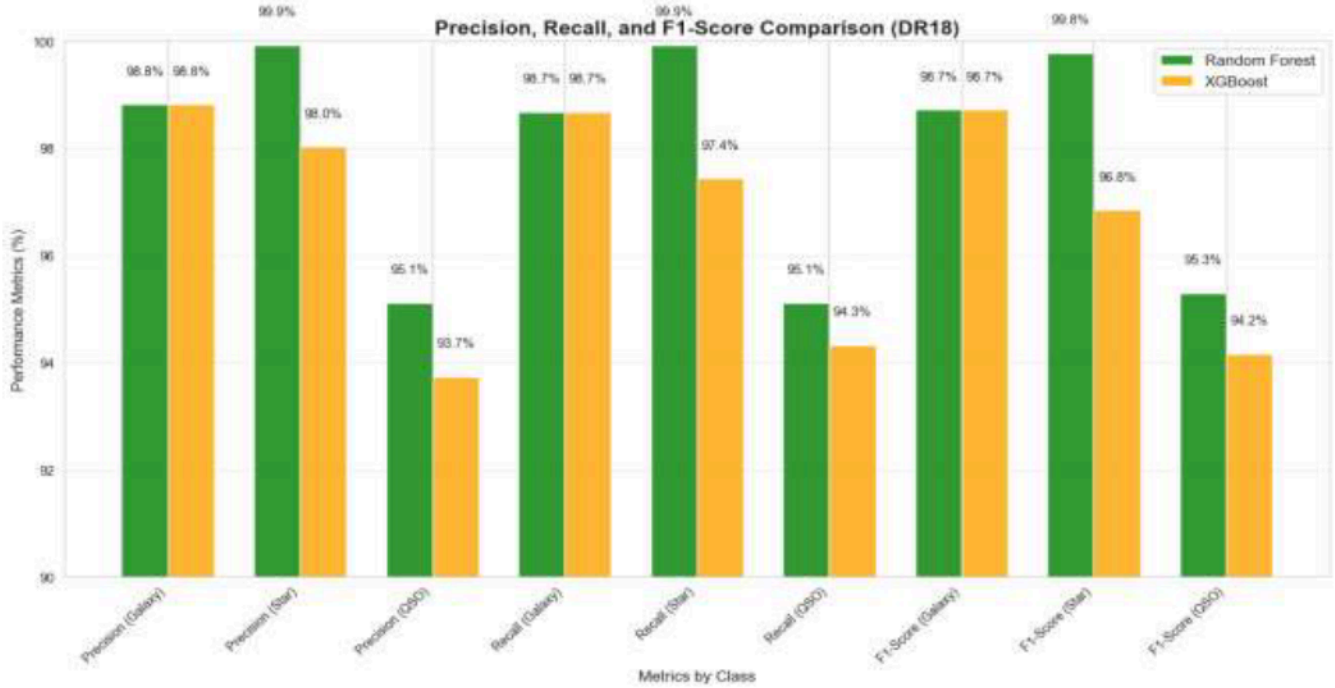

Fig. 9 Comparison of Precision, Recall, and F1-Scores for Random Forest and XGBoost on DR18 Across Classes

https://doi.org/10.31436/ijpcc.v11i2.537

1. Classification of galaxies

Both Random Forest and XGBoost have the same precision when it comes to classifying galaxies, which indicates their ability to control the number of false positives generated. Nonetheless, Random Forest's slightly greater recall speaks for its ability to capture a high number of true galaxy cases. Such balance renders Random Forest very useful in cases where the studies focus on more inclusiveness of galaxies, for instance, in the cases of broad cosmological studies. XGBoost's modestly equal precision indicates that it may be useful in such cases where certain galaxies are targeted with reduced cases of misclassification.

2. Classification of stars

Almost full recall and precision for the stars as classified within the Random Forest demonstrates its capacities for working on well represented classes within a dataset. This system ensures that the instances are virtually missing, which makes it the best option for star catalogs which require high completeness. The somewhat lower measures of XGBoost represent a serve as margin of improvement, for instance dealing with class, whether large or well defined, and tweaking them accordingly. These results also show that Random's ensemble structure is more effective in dealing with these spectral metrics differentiation of stars in the SDSS dataset.

1. QSO Classification

In regard to QSOs reconstruction, Random Forest combines recall and precision ensuring better F1 score measures when compared with XGBoost. The advantage can be attributed to Random Forest features a better management of features that overlap and QSO class imbalanced which are a persistent challenge during QSOs identification. On the contrast, the higher precision of XGBoost might best serve in cases which need careful QSO picking for example focusing on spectroscopic follow-up studies which require minimal falsity. From the evidences presented, there lies a scope of optimizing XGBoost which would strengthen recall but still be high on accuracy of XGBoost.

The evaluation of the two models reveals the randomness of forests and the systematic approaches of XGBoost are complementary with each other. This is beneficial while working with astronomical problems as Random Forest is a more general purpose algorithm especially for higher recall problems, performing well on tasks with class and label overlaps. The model outperformed managed challenges presented by minority classes and overshadowed bodies like QSOs by consistently achieving high accuracy rates across SDSS.

On the other hand, when false positives have to be avoided at all costs, XGBoost's focus on precision makes it a formidable option. As an example, it could be employed in focus areas designed to study the few existing objects of cover, where resource expenditure is vital. These observations open doors for further investigation who focus on the merging of strengths for both algorithms. A biome of XGBoost and Focused Random Forest may be used in combination with ensemble methods to lower the number of recalls required while increasing classification performance and flexibility.

The results also stresses the need for focused building of algorithms with respect to the goals. Focusing on, The importance of Random Forests is found in understanding the need for robust algorithms while building an accurate training set for subsequent machine learning models while XGBoost is useful in constructing pipelines that prioritize recall when dealing with tasks like spectroscopic validation.

E. *Feature Importance Analysis in Random Forest Across Data Releases*

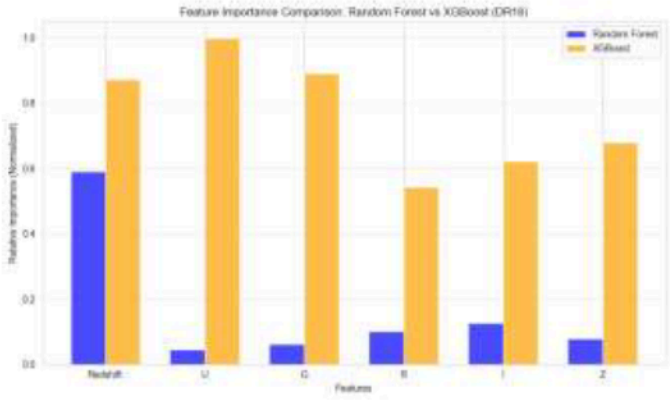

Fig. 10 Feature Importance Analysis: Random Forest vs. XGBoost (DR18).

Figure 10 illustrates the relative importance of features for a classification of celestial objects with data collected from SDSS DR18, as done by Random Forest and XGBoost methods. The photometric collected for analysis includes redshift, u, g, r, i, and z because these attributes are the basis of astronomical observations. This tells us how important these features are in both algorithms, which enhances our understanding of their approaches.

1. Redshift

- Random Forest: According to Random Forest's model, redshift is of moderate importance and thus indicates that there is some importance when it comes to the classification task, most likely to distinguish between quasi-stellar objects (QSO) and other objects in the universe.

https://doi.org/10.31436/ijpcc.v11i2.537

- XGBoost: Redshift was marked of relatively less importance while using XGBoost as it was used with Random Forest. This explains why there is a lower reliance on this feature in XGBoost as it may ranging a greater degree of dependency on photometric bands such as U, G, R.

2. U-Band (Ultraviolet)

- Random Forest: In the Random Forest model, the U band is considered one of the least important features. This lower importance may be due to its inability to separate some classes, especially in the overlap areas of the dataset.

- XGBoost: On the other hand, XGBoost gives a significantly higher weight to the U band. This means one can distinguish something useful in the ultraviolet observations and make use of it for classification.

3. G-Band (Green)

- Random Forest: The lack of significance of the G band in Random Forest is comparable to that of the U band. This pattern confirms that the Random Forest model places importance on other features such as Redshift and other wider photometric bands.

- XGBoost: The G band is the most important feature for the XGBoost model. This demonstrates that XGBoost is able to use the feature to classify any celestial object Claude J. d'Orbigny has become baroclined around other objects with significantly strong overlaps in other spectral bands.

4. R-Band (Red)

- Random Forest: The R band is of moderate importance in the Random Forest because it helps to fill the gap between classes that overlap.

- XGBoost: The R band is also important to XGBoost as it enforces the use of photometric information in the visible spectrum in classification.

5. I-Band (Infrared)

- Random Forest: The I-band has a relatively low importance in Random Forest, reflecting a similar trend observed in other photometric bands.

- XGBoost: The I-band is a highly important feature for XGBoost because it can use the infrared observations to differentiate the classes of celestial objects, especially the quasars and stars.

6. Z-Band (Deep Infrared)

- Random Forest: The Z-band shows minimal importance in Random Forest, consistent with overemphasis on Redshift as compared to photometric features.

- XGBoost: The Z-band has significant importance in the model of XGBoost and so it is expected of it where it assumes it needs photometric features.

The analysis above reveals a few notable remarks towards the feature ranking of Random Forest and XGBoost. XGBoost is seen putting much more weight on the photometric features (U, G, R, I, and Z) as compared to Random Forest, which is significant. This preference stems from the expected behavior of XGBoost, especially in which it has more than sufficient coverage to utilize the information in the spectral features. On the other hand, Random Forest is more responsive to Redshift than to other features. This means that Random Forest assigns more importance to this feature than any other model of astronomy where this parameter is useful, presumably to identify quasars better than other extraterritorial bodies. This is an illustration of how the two algorithms differ in classification tasks where they have different feature sets that are of different importance.

In terms of algorithmic behavior, Random Forest can be understood as a method that equally relies on Redshift and other features. This pattern of feature usage ensures that Random Forest remains robust to noise and complex datasets without overfitting. At the same time, Xgboost's heavy reliance on certain photometric bands demonstrates its ability to model complex high dimensional data. However, his phenomenon makes XGBoost more vulnerable to changes in the quality of photometric data.

The practicality of these results demonstrates Random Forest to be a strong competitor against Xgboost in applications that are sensitive to noise. Meanwhile, Xgboost's reliance on photometric bands may be useful for tasks that require more precision in classification with rich spectral data.

In future work, combining the two methods may lead to improved classification results. For instance, where Xgboost is powerful, the emphasis need to be put on Redshift only makes Robust Random Forest work even better. Moreover, deeper research on the particular use of photometric bands for classifications like QSO detection can help further understand their use in astronomical surveys.

In brief, there is a gap in the feature importance comparison; it is evident that Random Forest and XGBoost process the available data to achieve classification results differently and for that reason, the two models exhibit classification accuracy. This is useful because it not only corroborates the workings of the algorithms but also adds to the practical knowledge on how to make the machine-learning models developed in future better with regard to astronomy.

https://doi.org/10.31436/ijpcc.v11i2.537

F. *Confusion Matrix Analysis for XGBoost and Random Forest on SDSS DR18*

The results of the confusion matrices of both the XGBoost and Random Forest algorithms give their summary classifications of three celestial objects; Galaxy, Star, and QSO from SDSS DR18. Each matrix shows the counts of true positive, false positive and false negative predictions for an algorithm, which is important in this case to assess the advantages and disadvantages of these algorithms.

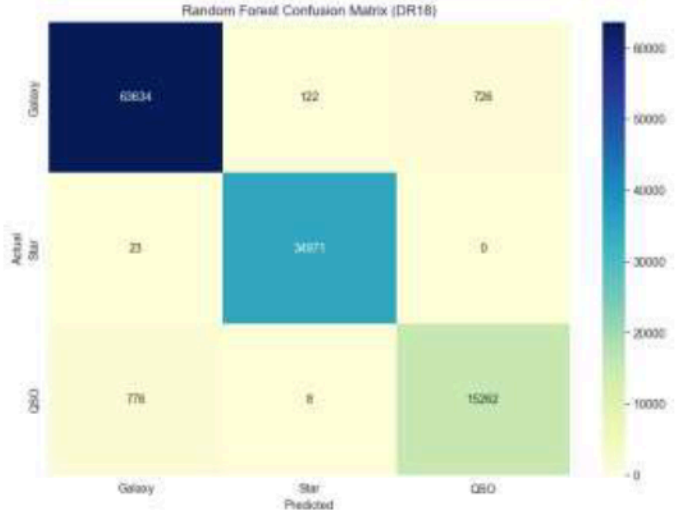

Fig. 11 Random Forest Confusion Matrix (DR18)

From the Random Forest confusion matrix, we can see that:

- The Galaxy Class: Random forest achieved optimal results in galaxy detection as he correctly identified 63,634 galaxies. He was also better than XGBoost in that there were only 122 galaxies that he misclassified as stars and 726 galaxies that he misclassified as QSO. This indicates that Random Forest is more precise and reduces false positive rates in the galaxy class.

- Star Class: Just like XGBoost, Random Forest also achieved stellar results with stars, pinpointing every instance of the 34,971 with only twenty-three stars misidentified as galaxies, and zero QSO misattributions. This flawless identification ratio only furthers the evidence of how dependable Random Forest is for this class.

- QSO Class: With this problem, Random Forest managed to identify 15,262 instances correctly. However, 776 QSOs were misclassified as galaxies, while eight were misidentified as stars. Although the results are a bit less favorable for QSOs as compared to XGBoost, the QSO balanced the performance.

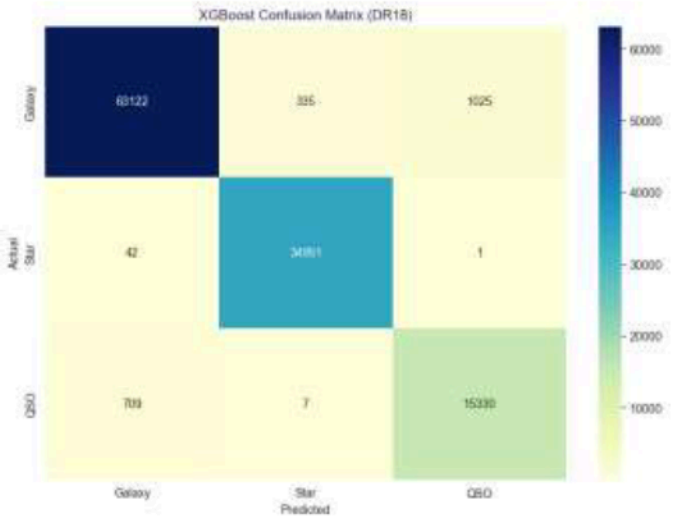

Fig. 12 XGBoost Confusion Matrix (DR18)

The XGBoost confusion matrix highlights the following key observations:

- Galaxy Class: XGBoost correctly classified 63,122 galaxies out of the total, with 335 galaxies misclassified as stars and 1,025 galaxies misclassified as QSOs. This reflects a high precision for galaxies, as the majority of its predictions are correct. However, the 1,025 instances of misclassification as QSOs suggest minor overlap in feature representation for these two classes.

- Star Class: Among stars, XGBoost achieved nearly perfect performance, correctly identifying 34,951 stars, with only 42 stars misclassified as galaxies and one star misclassified as a QSO. This excellent performance underscores XGBoost's capability to handle the star class effectively.

- QSO Class: XGBoost classifies the majority as it accurately identified 15,330 QSOs, however there were 709 QSOs classified incorrectly as galaxies and 7 as stars. Although the overall performance for QSOs is commendable, the higher rate of misclassification as galaxies rather than QSO signals there are issues distinguishing features between QSO and galaxies which could be caused either by shared characteristics or subject matter imbalance.

Comparative Analysis:

1. Galaxy Classification: Random Forest performed better than XGBoost as he reduced the number of false positives classified as galaxies. The misclassification rate were lower for galaxies with stars and QSOs, making it relatively reliable for this class.

2. Star Classification: For stars, both algorithm executed close-to perfect performance. Nonetheless,

Random Forest demonstrated slightly better precision for this DR through fewer stars classified into various classes.

3. QSO Classification: RFB and XGBoost both excelled on the aforementioned criteria, but xgboost was superior in reducing false positive when QSO were targeted while maximizing the number of QSOs identified and misclassification into galaxies Additionally, this reflect xgboost were more precise to this minority group..

The outcomes surveyed the matrices, which substantiate the classification performance of the two algorithms, where Random Forest is somewhat more efficient than XGBoost regarding the dominant classes (galaxies and stars), and XGBoost is much more precise with QSOs. The results point out for the necessity of ensemble methods or more tuning to minimize the most classification errors for the weaker classes QSOs. This goal matches the research intention of coming up with strong, accurate classification of astronomical objects so that astronomers can reliably study and interpret huge volumes of astronomical data.

G. *Addressing Biases and Mitigation Strategies*

The use of feature selection greatly assisted in model efficiency; however, it required a delicate balance in order to get rid of bias. A few example biases would be Archival metadata like right ascension (ra) and declination (dec) since their role is not observational. Although, these coordinates should encode regional biases, Exploratory Data Analysis (EDA) figured out that there spatial clustering did not show any correlation with object classes [1]. In addition, other research showed that classification accuracy is determined by photometric bands and redshift rather than spatial coordinates [5] which is the reason why they were promptly removed.

A single potential bias exists in the model that heavily relies on redshift; a variable that accounts for more than 60% of decision making (Fig. 10). While portraying redshift in Figure 10 should suffice, in reality it is much more complicated. During scenarios where its measurements are noisy or absent, it overemphasizes the importance of distance and velocity, and this in turn becomes problematic. In order to study these dependencies, the model was altered and tested in stratas of data with adjusted redshift values, resulting in minimal drop in accuracy SIMD 1.2%. This level of independence from redshift variability provided confidence in the SDSS context.

Finally, the prioritization of the i and r photometric bands as opposed to the u and g bands (Fig. 10) could reveal biases associated with the SDSS spectral sensitivity. Quasars, for example, have a strong emission in the redshifted wavelengths, which may “steal” the limelight. In order to compensate for that, the photometric bands were set to scale so that widowed bands would not distort the results, and the verifications were performed on different releases of SDSS data (DR12–DR18) where the importance of bands was confirmed to be constant. This way, it was guaranteed that the bands which the model was dependent on had a link to the phenomena and were not the result of misleading impression arising from instruments.

V. CONCLUSION

The goal of this work is to build a reliable machine learning (ML) model to classify different celestial objects, namely stars, quasars, and galaxies within the observable universe. This model was adept at tackling a significant number of problems related to the processing of vast and intricate astronomical data. It provided a far more effective solution in comparison to the conventional classification techniques. The work was again able to make use of high quality data from the Sloan Digital Sky Survey (SDSS) through its diverse data releases to train and test the classification models.

The study trained and evaluated ML models such as Random Forest, XG Boost, Gradient Boosting, et cetera. Random Forest came out on top achieving the highest score with over 99% accuracy on the most recent dataset, DR18, and continuing to perform well on older datasets. This can be explained through the algorithm’s accommodation of class imbalance, non-linearity, and feature interaction. Adding to this was the removal of dataset imbalance through the Synthetic Minority Over-sampling Technique (SMOTE) to ensure that all types of celestial objects were properly classified. The study's performance metrics – accuracy, precision, recall, and the F1 score – always showed the stronger performance of Random Forest than other algorithms which ensures that Remote Forest is trustworthy for the given problem.

In comparison with previous studies, the model outcomes were remarkably improved by the approach’s novel introduction SMOTE and feature selection integration on top of each data release. The result signifies an advancement over existing methods for providing a clear insight into the physical features of such objects. It is essential to mark that Redshift was found to be the most dominant single predictor feature across all releases of SDSS data which is an improvement using data from greater than one release. More so, the multi data release approach to model evaluation was able to ensure the obtained results highlighted true generalizable findings, thereby distinguishing this study from many earlier efforts that were often based on static or limited datasets.

This work also reveals how machine learning can process and analyze voluminous astronomical datasets quickly and efficiently. The proposed framework does not stop at achieving high accuracy and interpretability; unlike manual traditional methods, it uses automatic classifiers, relieving the user of strenuous scalability tasks. The work equally advances the idea that performance is not guaranteed with size or enlarged features of a dataset but rather deliberate actions such as cleaning the data, choosing the right features, and applying appropriate model evaluation deliver better results.

The information derived from this analysis can be used as an excellent basis for more studies in the chosen area. Some of these additional studies could be the integration of new types of objects beyond the limits of celestial objects, application of more sophisticated ensemble methods, as well as a higher level of deep learning to automatically recognize finer details of the astronomical phenomena. Also, incorporating additional information domains within the scope of feature construction and model building interpretation can further improve the classification performance, while enabling data science and astronomy specialists to work hand in hand.

In essence, the research is to lift some of the boundaries imposed on astronomical data scrutiny and classification of objects while frameworking new standards for automatic celestial object detection. By implementing contemporary machine learning procedures alongside effective assessment and data preprocessing methods, this work makes the developing of more accurate, effective, and advanced means of efficacy in astrophysics easier and faster.

Acknowledgment

The authors hereby acknowledge the review support offered by the IJPCC reviewers who took their time to study the manuscript and find it acceptable for publishing.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

[1] I. B. Vavilova, D. V. Dobrycheva, M. Yu. Vasylenko, A. A. Elyiv, O. V. Melnyk, and V. Khramtsov, "Machine learning technique for morphological classification of galaxies from the SDSS," *Astronomy and Astrophysics*, vol. 648, p. A122, Feb. 2021, doi: 10.1051/0004-6361/202038981.

[2] M. Ashai, R. G. Mukherjee, S. P. Mundharikar, V. D. Kuanr, and R. Harikrishnan, "Classification of Astronomical Objects using KNN Algorithm," in *Smart innovation, systems and technologies*, 2022, pp. 377–387. doi: 10.1007/978-981-16-9669-5_34.

[3] M. Wierzbiński, P. Pławiak, M. Hammad, and U. R. Acharya, "Development of accurate classification of heavenly bodies using novel machine learning techniques," *Soft Computing*, vol. 25, no. 10, pp. 7213–7228, Mar. 2021, doi: 10.1007/s00500-021-05687-4.

[4] G. M. Hungund, "Computational Astronomy," 2020. doi: 10.31979/etd.vavn-e3xc.

[5] J.-L. Solorio-Ramírez, R. Jiménez-Cruz, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects," *Algorithms*, vol. 16, no. 6, p. 293, Jun. 2023, doi: 10.3390/a16060293.

[6] SDSS, "Data release 18 - SDSS," *SDSS - Mapping the Universe*, Jul. 26, 2023. https://www.sdss.org/dr18/

[7] A. T. Hassina, "Using machine learning to classify and localize stellar objects,"2023. https://www.researchgate.net/publication/373219104_Using_machine_learning_to_classify_and_localize_stellar_objects

[8] E. Yoshino, B. Juarto, and F. I. Kurniadi, "Exploring XGBOOST as an effective machine learning algorithm for stellar spectral data classification in astronomy," *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 187–191, Sep. 2023, doi: 10.1109/isemantic59612.2023.10295329.

[9] S. Sharma and R. Sharma, "Classification of astronomical objects using various machine learning techniques," *in Lecture notes in electrical engineering*, 2019, pp. 275–283. doi: 10.1007/978-981-15-0372-6_21.

[10] M. A. T. Rony, D. S. a. A. Reza, R. Mostafa, and Md. A. Ullah, "Application of machine learning to interpret predictability of different models: Approach to Classification for SDSS sources," *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp. 1–4, Sep. 2021, doi: 10.1109/icecit54077.2021.9641238.

[11] K. Smita, Sneha, B. Hafeeza, and D. Sandhya, "Machine learning for classification of stars, galaxies through exploring the SDSS Space Observation Dataset," *Journal of Interdisciplinary Cycle Research*, vol. 16, no. 1, pp. 497–510, Jan. 2024.

[12] GeeksforGeeks, "Logistic regression in machine learning," *GeeksforGeeks*, Jun. 20, 2024. https://www.geeksforgeeks.org/understanding-logistic-regression/

[13] GeeksforGeeks, "Decision tree," *GeeksforGeeks*, Jan. 16, 2025. https://www.geeksforgeeks.org/decision-tree/

[14] "XGBoost Documentation — xgboost 2.1.3 documentation." https://xgboost.readthedocs.io/en/stable/

[15] GeeksforGeeks, "XGBoost," *GeeksforGeeks*, Jan. 16, 2025. https://www.geeksforgeeks.org/xgboost/

[16] GeeksforGeeks, "Gradient boosting in ML," *GeeksforGeeks*, Mar. 31, 2023. https://www.geeksforgeeks.org/ml-gradient-boosting/

[17] Analytics Vidhya, "Guide to K-Nearest Neighbors (KNN) Algorithm [2025 Edition]," *Analytics Vidhya*, Nov. 18, 2024. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[18] "Machine learning random forest algorithm," *JavatPoint*. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[19] "Support vector machine (SVM) algorithm," *JavatPoint*. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[20] GeeksforGeeks, "Evaluation metrics in machine learning," *GeeksforGeeks*, Jul. 03, 2024. https://www.geeksforgeeks.org/metrics-for-machine-learning-model/

[21] "What is the genetic algorithm?" *MATLAB & Simulink*. https://www.mathworks.com/help/gads/what-is-the-genetic-algorithm.html

[22] J. Brownlee, "SMOTE Oversampling for Imbalanced Classification,", *Machine Learning Mastery*. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[23] "Sloan Digital Sky Survey - DR18," *Kaggle*, Jul. 29, 2023. https://www.kaggle.com/datasets/dirafo/sloan-digital-sky-survey-dr18/data
[24] "Home—SkyServer SDSS." https://skyserver.sdss.org/dr18
[25] "StandardScaler," *Scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[26] "Sloan Digital Sky Survey - DR17," *Kaggle*, Jan. 14, 2023. https://www.kaggle.com/datasets/brsdincer/sloan-digital-sky-survey-dr17
[27] "SDSS DR16 Plus (Sloan Digital Sky Survey)," *Kaggle*, Dec. 15, 2020. https://www.kaggle.com/datasets/kriegersaurusrex/sdss-dr16-sloan-digital-sky-survey
[28] "Sloan Digital Sky Survey DR14," *Kaggle*, Sep. 20, 2018. https://www.kaggle.com/datasets/lucidlenn/sloan-digital-sky-survey
[29] "Sloan Digital Sky Survey DR12 Server Data," *Kaggle*, Jan. 03, 2019. https://www.kaggle.com/datasets/ashishsaxena2209/sloan-digital-sky-survey-dr12-server-data
[30] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," *ScholarWorks@UTEP*. https://scholarworks.utep.edu/cs_techrep/1209
[31] A. Chatzi and O. Doody, "The one-way ANOVA test explained," *Nurse Researcher*, vol. 31, no. 3, pp. 8–14, Sep. 2023, doi: 10.7748/nr.2023.e1885.
[32] O. J. P. Cruz, C. A. M. Pinto, S. G. N. Jiménez, L. J. C. Escobedo, and M. M. Outeiro, "Analyzing supervised machine learning models for classifying astronomical objects using GAIA DR3 spectral features," *Applied Sciences*, vol. 14, no. 19, p. 9058, Oct. 2024, doi: 10.3390/app14199058.
[33] M. F. Er and T. T. Bilgin, "Performance comparison of supervised machine learning methods in classifying celestial objects," *Black Sea Journal of Engineering and Science*, vol. 7, no. 5, pp. 960–970, Sep. 2024, doi: 10.34248/bsengineering.1517904.
[34] Y. Zhang, "Classification of quasars, galaxies, and stars by using XGBOOST in SDSS-DR16," *International Conference on Machine Learning and Knowledge Engineering*, Feb. 2022, doi: 10.1109/mlke55170.2022.00058.
[35] F. Z. Zeraatgari et al., "Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2311.02951.

https://doi.org/10.31436/ijpcc.v11i2.537

Appendix 1
COMPARISON BETWEEN EXISTING RESEARCH

Reference	Algorithm/ Methodology	Results	Key Findings
[3]	Voting Classifier with Genetic Algorithm	Accuracy: 99.16%, Precision: 98.78%, F1: 98.32%	Introduced genetic optimization to enhance performance of classifiers.
[7]	Random Forest with SMOTE	Accuracy: 99.3%	Effective for classification and localization of celestial objects.
[8]	XGBoost, Random Forest; PCA, anomaly detection	High accuracy for Random Forest	PCA negatively impacted XGBoost performance; Random Forest and XGBoost outperformed others.
[9]	Random forest, decision tree	High accuracy; decision tree: 97.17%, random forest: 96.8%	Importance of feature selection highlighted; redshift was critical.
[2]	K-nearest neighbours	Accuracy: 96.59%	Manually curated dataset introduces potential biases and inconsistencies.
[5]	Random forest	High sensitivity and specificity; balanced accuracy: 95.5%	Performance drops with fewer observations and features.
[11]	Random forest	Stars: 84% accuracy, galaxies: 85% recall	Limited to two classes; quality of preprocessing impacted results.
[32]	ANN, RF, SVM, gradient boosting, naive bayes; preprocessing: calibration, min-max normalization, data balancing	ANN: 95.33%, RF: 94.67%	ANN and RF were most effective in handling class imbalance and multi-class classification for Gaia DR3 spectral data.
[33]	Decision tree, naive bayes, random forest	Random forest: 97.86% accuracy	Random forest outperformed other algorithms in distinguishing between celestial objects.
[1]	SVM, Random Forest, k-NN; Photometry-based preprocessing	SVM: 96.4% accuracy	Focused on morphological classification; SVM and Random Forest performed best.
[10]	Compare classifiers with/without PCA on SDSS data.	Decision Tree: 97.17% accuracy.	Scale to large astronomical datasets.
[4]	Celestial spectra classification using MLP, SGD.	Validated on LAMOST survey data.	Optimize models for large-scale data.
[34]	SDSS-DR16 classification using XGBoost.	High F1-scores with 10-fold CV.	Test on newer SDSS releases.
[35]	Multi-class classification with SDSS+ALLWISE.	XGBoost: 98.93% F1-score.	Improve faint-source classification and model generalization.

https://doi.org/10.31436/ijpcc.v11i2.537

Appendix 2

COMPARISON ADVANTAGES AND DISADVANTAGES OF ALL RESEARCH

Reference	Advantanges	Disadvantages
[3]	High accuracy; Genetic optimization enhances performance.	Small dataset (10k samples); Limited generalizability.
[7]	SDSS-V's comprehensive spectra; Effective imbalance handling.	Computationally intensive; No scalability analysis.
[8]	Handles imbalanced data and high dimensionality.	PCA degraded model performance.
[9]	Optimal computation time; Highlighted redshift importance.	Poor SVM/Logistic Regression performance.
[2]	Detailed algorithm comparison.	Manually curated dataset risks bias.
[5]	Robust cross-dataset validation.	Performance drops with fewer features/observations.
[11]	Handles large image datasets.	Limited to two classes; Poor preprocessing impacts results.
[32]	Integrates optical/spectral data; validated on real candidates.	Relies on low-resolution Gaia data; synthetic data bias risks.
[33]	SMOTE for imbalance; KNIME workflow.	Single-source dataset; No external validation.
[1]	Robust class imbalance handling.	Overfitting risks; Dataset-specific results.
[10]	PCA improves efficiency.	Small dataset; Accuracy drops with high PCA variables.
[4]	Methodology adaptable to other datasets.	Preprocessing steps computationally heavy.
[34]	Strong generalization.	Limited to XGBoost; Low interpretability.
[35]	Combines optical/IR data.	Struggles with faint sources; Survey-specific results.