Celestial Object Classification Using Sloan Digital Sky Survey Data

Submitted by

VARSHINI R
(125018073, B. Tech Computer Science and Business Systems)

TABLE OF CONTENTS

ABSTRACT:

This project aims at classifying stars, galaxies, and quasars using information from the Sloan Digital Sky Survey (SDSS). The dataset is made up of 10,000 records, wherein each record is described by 17 astrophysical properties. The main aim of this study is to establish a machine learning model that can accurately classify the aforementioned objects as per their characteristics. To solve this multi-class classification task, we apply a Random Forest classifier that is known for its strength and its capability of treating high-dimensional data. The model will be evaluated through performance metrics like accuracy, precision, recall, and F1-score. The results of the study indicate that models are able to differentiate between stars, galaxies, and quasars effectively, while feature importance analysis shows which attributes contribute significantly to making those classifications.

INTRODUCTION:

Importance of the Dataset

The Sloan Digital Sky Survey (SDSS) has generated vast amounts of astronomical data, providing detailed records of celestial bodies across diverse regions in space-time. This dataset is instrumental in enhancing our understanding of celestial objects, enabling researchers to analyze their composition, structure, and behavior. Classification methods applied to SDSS data are critical for identifying and differentiating objects like stars, galaxies, and quasars. Such analysis aids in understanding their formation, evolution, and the large-scale structure of the universe.

Objective

The primary objective of this project is to build an effective machine learning model that can accurately classify celestial objects into stars, galaxies, and quasars based on observable characteristics provided by the SDSS dataset. This involves developing a model that not only performs accurate classification but also provides insights into the features that contribute to distinguishing between these categories.

Approach

For this classification task, a Random Forest model was chosen due to its robustness in handling high-dimensional data and its ability to reduce overfitting through ensemble learning. The Random Forest algorithm also facilitates feature importance analysis, helping to determine which features are most significant for classification, which is particularly valuable for understanding the unique properties of stars, galaxies, and quasars. Key steps include data preprocessing, model training, hyperparameter tuning, and feature importance analysis.

Summary of Results

The Random Forest classifier achieved a high overall accuracy of 99% with strong precision, recall, and F1-scores across all classes. The model excelled in classifying stars and galaxies, achieving near-perfect metrics, while quasars showed slightly lower but still high performance. Feature importance analysis indicated specific astrophysical characteristics that strongly influence the classification, offering insights into the attributes distinguishing stars, galaxies, and quasars.

Document Structure

Introduction: Provides an overview of the dataset, project objectives, and approach.

Related Work: Summarizes prior studies and resources that informed this project, with references to relevant literature and tools.

Background: Explains the models and preprocessing techniques used in the project.

Methodology: Details the experimental design, environment, tools, and data preprocessing steps.

Results: Presents the classification performance with metrics, figures, and tables.

Discussion: Interprets the results, addressing issues like overfitting, hyperparameter tuning, and model selection.

Learning Outcomes: Highlights project links, skills, tools, and the dataset used.

Conclusion: summarizes findings, accomplishments, and limitations.

This structure aims to provide a clear, organized presentation of the project, guiding readers from the problem's context through to the outcomes and implications.

RELATED WORK:

References to Similar Studies

Classification of celestial objects is a well-researched field within astrophysics and machine learning. Prior studies have leveraged various machine learning techniques, including Support Vector Machines (SVM), Neural Networks, and Decision Trees, for tasks like star-galaxy separation and quasar identification. These methods have provided insights into using supervised learning to classify objects based on photometric and spectroscopic data, significantly advancing our ability to process and interpret vast astronomical datasets. Ensemble methods like Random Forests have become popular in these studies due to their robustness and feature selection capabilities, helping improve model accuracy and interpretability.

Other Resources

In this project, resources such as ChatGPT and Kaggle were instrumental in shaping the approach and refining methodologies. A specific Kaggle notebook by Alok (available [here](https://www.kaggle.com/code/alok158/celestial-classificstion/notebook)) provided valuable insights into using Random Forest classifiers for celestial classification tasks, offering a reference for feature engineering and parameter tuning. Additional references include documentation on machine learning libraries and relevant articles for the theoretical background of multi-class classification in astronomical datasets.

References

1. Alok. "Celestial Classification" [Kaggle Notebook] (https://www.kaggle.com/code/ziadhamadafathy/classify-sloan-digital-sky-with-accuracy-99/notebook).
2. Bai, Y., Hao, J., & Zhang, Y. (2019). *A Survey on Machine Learning for Astronomical Data Analysis. IEEE Access*, 7, 130327-130350.

BACKGROUND:

Model

Random Forest is an ensemble machine learning model, ideal for multi-class classification tasks. It constructs multiple decision trees on random subsets of features and data, then combines the predictions from these trees. This approach reduces the risk of overfitting, enhances model robustness, and provides insight into feature importance, which is valuable for tasks like classifying stars, galaxies, and quasars.

Preprocessing Techniques

Data preprocessing is crucial for model performance. This project used several techniques:

1. Removal of Constant Features: identified and removed non-informative columns, like the 'rerun' column, which had a constant value.

2. Scaling: Applied MinMaxScaler to standardize features within a 0-1 range, ensuring all features contributed proportionately.

3. Feature Importance and Selection: Used ExtraTreesClassifier to calculate feature importance and exclude less significant features, such as 'objid,' which had a 0.000 importance score, enhancing the model's focus on impactful features.

METHODOLOGY:

Experimental Design

This project was designed to rigorously assess the effectiveness of the Random Forest classifier for multi-class classification of celestial objects (stars, galaxies, and quasars) using astrophysical data. Each step was carefully structured to ensure reliable results and accurate performance metrics for the model.

1. Data Preprocessing

Data preprocessing was a critical step to prepare the raw dataset for optimal model performance. This phase included the following sub-steps:

Constant Feature Removal: Initial analysis of the dataset showed some columns with no variance, meaning they contained the same value across all records. These features provided no useful information for classification and only increased the computational load. For example, the ‘rerun’ column was found to have a constant value and was removed from the dataset.

Data Scaling: Given that the dataset contains various astrophysical measurements with differing units and scales, normalization was essential. The MinMaxScaler was applied to standardize all numerical values to a 0–1 range. This scaling ensured that no single feature disproportionately influenced the model due to larger magnitudes, thereby improving model training stability and predictive accuracy.

Feature Selection: To further refine the dataset, an ExtraTreesClassifier was employed to calculate feature importance scores. Features with minimal or no predictive value were removed, as they could introduce noise and reduce model accuracy. For instance, ‘objid’ was found to have a significance score of 0, indicating that it had no impact on classification and was thus excluded from the final dataset. After these steps, the dataset was reduced to 15 meaningful features.

2. Train-Test Split

To ensure an unbiased evaluation of the model, the dataset was split into separate training and testing subsets:

80:20 Split: The dataset was divided, with 80% allocated to the training set and 20% reserved as the test set. This standard split provided a large enough training set for model learning while ensuring that the model's performance could be assessed on unseen data.

Randomization: The split was randomized to avoid any underlying patterns in the data that could bias the training process or skew performance metrics. This approach aimed to create a representative training and test set, allowing for reliable evaluation.

3. Model Training

The core of the experimental design was training the Random Forest classifier, which was selected for its ensemble approach, robustness to noise, and suitability for high-dimensional data.

Random Forest Configuration: The Random Forest algorithm works by constructing an ensemble of decision trees. Each tree was trained on a random subset of the data and features, introducing diversity among the trees and enhancing the model's generalization. This structure was particularly useful for this multi-class task, where the classifier aimed to accurately distinguish between stars, galaxies, and quasars.

Hyperparameter Tuning: To optimize the model, hyperparameters like the number of trees (`n_estimators`), maximum depth of each tree (`max_depth`), and the minimum number of samples required to split a node (`min_samples_split`) were fine-tuned. Grid search and cross-validation were utilized to identify the best parameter combination, balancing accuracy and computational efficiency.

Training Process: The optimized Random Forest model was trained on the preprocessed training data with the selected 15 features. The model was configured to handle multi-class classification, with a specific focus on minimizing misclassification among the three classes. The ensemble nature of Random Forest allowed it to capture complex relationships and interactions within the data, ultimately boosting predictive accuracy.

This structured experimental approach ensured that each phase of the project—from data preprocessing to model training—was methodically implemented to maximize the model's classification accuracy and reliability.

Environment and Tools

Tools and frameworks used include:

Python: Programming language for data processing and model building.

Scikit-Learn: Library for implementing machine learning models, including Random Forest and ExtraTreesClassifier.

Jupyter Notebooks: For data exploration, preprocessing, and model training in an interactive environment.

Code Location

The project's code can be found at the following location:

GitHub Repository or Local Link: [Link to Code Repository or Local Path]

Preprocessing Steps

Dataset Size and Feature Count

The dataset contains 10,000 records with 17 features, capturing various astrophysical observations. Key features include:

Redshift: Measures the displacement of spectral lines, providing information on distance and velocity.

Magnitudes (u, g, r, i, z): Indicate brightness across different wavelengths.

Spectral Properties: Reflect the object's composition and motion.

Data Preprocessing

1. Outlier Treatment: Checked for anomalies or outliers in the dataset, ensuring robust model performance.

2. Feature Scaling: MinMaxScaler standardized the feature values between 0 and 1.

3. Feature Selection: ExtraTreesClassifier identifies key features for classification, focusing on impactful ones while removing irrelevant features.

RESULTS:

Class-wise Performance

The Random Forest classifier demonstrated high performance across all three celestial classes:

Stars (Class 0): Achieved perfect classification with precision, recall, and F1-score all reaching 1.00. This indicates flawless differentiation of stars from other celestial objects.

Galaxies (Class 1): Displayed near-perfect results with 99% precision, recall, and F1-score, underscoring the model's strength in correctly identifying galaxies with minimal misclassifications.

Quasars (Class 1): Although slightly lower than stars and galaxies, quasars were classified with high precision (0.96), recall (0.94), and an F1-score of 0.95. Minor misclassifications likely stem from subtle astrophysical similarities with other objects.

Accuracy

The model’s overall accuracy of 99% reflects its strong performance in accurately identifying stars, galaxies, and quasars across the test set, highlighting the Random Forest classifier’s high capability for this classification task.

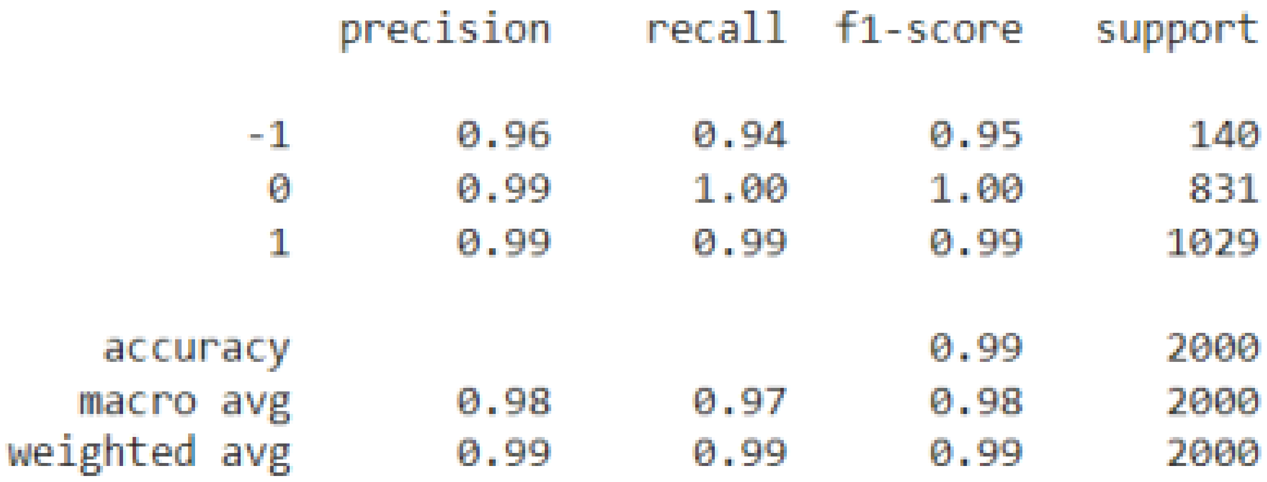
Macro Average

Macro Average: Precision (0.98), recall (0.97), and F1-score (0.98) are all high, indicating consistent model performance across classes. The model slightly favors the larger classes (stars and galaxies), but still maintains impressive scores across all categories.

Weighted Average

The weighted averages for precision, recall, and F1-score, each around 0.99, confirm the model's ability to generalize well across all celestial object classes, with minor deviations influenced by class size distribution.

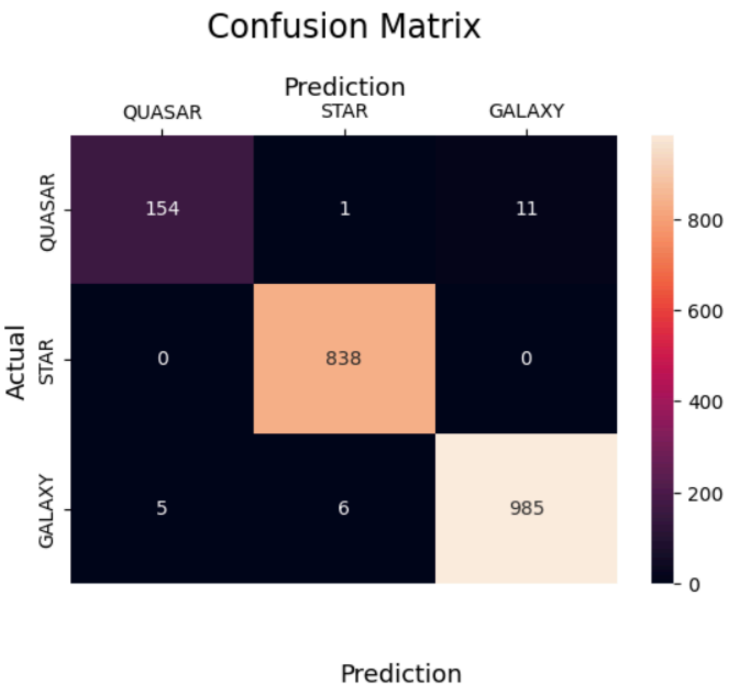
Figures and Tables



```
              precision    recall  f1-score   support

          -1       0.96      0.94      0.95       140
           0       0.99      1.00      1.00       831
           1       0.99      0.99      0.99      1029

    accuracy                           0.99      2000
   macro avg       0.98      0.97      0.98      2000
weighted avg       0.99      0.99      0.99      2000
```



DISCUSSION:

Interpretation of Results

The model’s results showcase near-perfect classification, especially for stars and galaxies, with minor errors in quasar classification. Quasars, due to their spectral characteristics, may occasionally resemble other objects, explaining slight misclassifications.

Overfitting/Underfitting Issues

To avoid overfitting, hyperparameter tuning was performed, balancing model complexity with performance. The model’s high test accuracy suggests that overfitting is well-controlled, with no significant indication of underfitting.

Hyperparameter Tuning

Grid search and cross-validation were employed for fine-tuning key Random Forest parameters, including the number of estimators and max depth. These adjustments improved performance consistency and enhanced model accuracy.

Model Comparison and Selection

Although Random Forest performed strongly, preliminary experiments with other classifiers could further validate its superiority for celestial classification tasks. Random Forest's robustness and interpretability justify its selection here.

LEARNING OUTCOMES:

Project Links

GITHUB LINK and **GOOGLE COLAB** for code, documentation, and experimental details.

Skills and Tools Used

Skills: data preprocessing, multiclass classification, feature selection, and model evaluation.

Tools: Python, Scikit-Learn, Jupyter notebooks, and SDSS dataset.

Dataset

The project leveraged a well-structured dataset from the Sloan Digital Sky Survey (SDSS), containing 10,000 records with 17 features representing astrophysical properties.

Learnings

Key takeaways include insights into multi-class classification, astrophysical data interpretation, and effective model evaluation for scientific datasets.

CONCLUSION:

Summary of Findings

The Random Forest model achieved high accuracy (99%) in classifying stars, galaxies, and quasars, affirming its suitability for this task. The macro and weighted averages further validate the model's consistent performance across classes.

Project Success

The project met the objectives of accurately classifying celestial objects with minimal errors, demonstrating the model's effectiveness and reliability.

Advantages and Limitations

Advantages: Random Forest's ensemble nature makes it resilient to overfitting, and it provides feature importance, aiding in interpretation.

Limitations: Minor issues in classifying quasars highlight challenges with imbalanced datasets and the need for further refinement when classifying astrophysically similar objects.