Stellar Classification Analysis

Group 10

Shruti Agarwal, SXA220122; Hao-Yu Chou, HXC230020

Ruocheng Jiang, RXJ220041; Hriday Muppidim, HXM220051

Surya Vamshi Sriperambudooru, SXS230148

The University of Texas at Dallas

6341.001.Applied Machine Learning

May 07, 2024

Stellar Classification Analysis

Abstract

This project proposes an innovative approach to advance the fundamental study of stellar classification within astronomy. Stellar classification plays a pivotal role in deciphering the composition, structure, and evolutionary paths of stars, galaxies, and quasars based on their unique spectral characteristics. By harnessing supervised learning techniques, our aim is to develop a sophisticated model capable of accurately categorizing these celestial entities.

To achieve our objective, we utilize a comprehensive dataset sourced from Kaggle, comprising a diverse array of observations collected by the Sloan Digital Sky Survey. Each observation is characterized by a wealth of features, including spectral data and class labels indicating whether it corresponds to a star, galaxy, or quasar. By leveraging the power of machine learning, we aspire to contribute to the advancement of astronomical research, enabling more accurate and efficient stellar classification and fostering new discoveries about the cosmos.

Data Set

The dataset used in this project is sourced from Kaggle (https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17). With 100000 observations of space taken by the Sloan digital sky survey. Every observation is described by 17 featured columns including one class column, which can be classified to be a star, a galaxy and quasar.

Methodology

This project utilizes exploratory data analysis (EDA) to provide insights into the dataset's quality, distribution, and inter-variable relationships. It enhances data understanding through renaming columns for better clarity, detecting missing or duplicated entries, and analyzing class distributions through multivariate analysis, where stellar data scatter plots with categorical classes visualizes astronomical data categorized into three classes, then correlation and distribution analyses further elucidate the relationships and characteristics of the data. Further, supervised and unsupervised models are implemented to accurately classify the data into different classes based on selected features.

Data Description

In astronomy, the classification of stars is based on their spectral properties, forming a core part of the broader scheme used to categorize galaxies, quasars, and stars. Initially, cataloging stars and mapping their positions helped astronomers realize that these stars form our galaxy. The discovery that the Andromeda nebula was a distinct galaxy from our own marked a pivotal moment, leading to the surveying of numerous other galaxies as more advanced telescopes were developed. This dataset focuses on classifying stars, galaxies, and quasars by examining their spectral features.

The data contains 100000 observations of space taken by the SDSS (Sloan Digital Sky Survey) with 18 attributes, with 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

Exploratory Data Analysis

Multivariate Analysis

Stellar Data Scatter Plot with Categorical Classes

The scatter plot visualizes astronomical data categorized into three classes—Galaxies (blue), Quasi-Stellar Objects or QSOs (orange), and Stars (green)—using right ascension angle on the x-axis and declination angle on the y-axis.

Stars concentrated mainly in discrete clusters, possibly indicating groupings based on their proximity or association with specific celestial features. Their spread is less extensive compared to galaxies, suggesting a more localized observation or inherent clustering in their spatial distribution.

Galaxies and QSOs are more widely distributed across the plot, with galaxies predominantly forming a large cluster in the central region, and QSOs interspersed within this cluster.

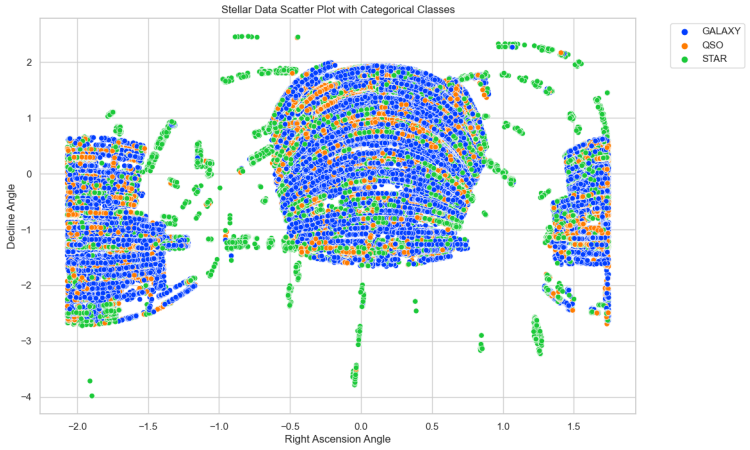

Figure 1

Stellar Data Scatter Plot with Categorical Classes (X-axis: right ascension angle, Y-axis: declination)

The second scatter plot displays galaxies, quasars, and stars against redshift and plate ID. Galaxies are marked in blue and mostly cluster at lower redshift values, suggesting they are closer to Earth or moving away more slowly. Quasars (orange) occupy higher redshifts, suggesting they are distant and fast-moving, aligning with their nature as bright, distant galactic cores. Stars (green) cluster near zero redshift, showing little to no motion away from Earth, and are uniformly distributed across all plate values, indicating a consistent inclusion in the survey.

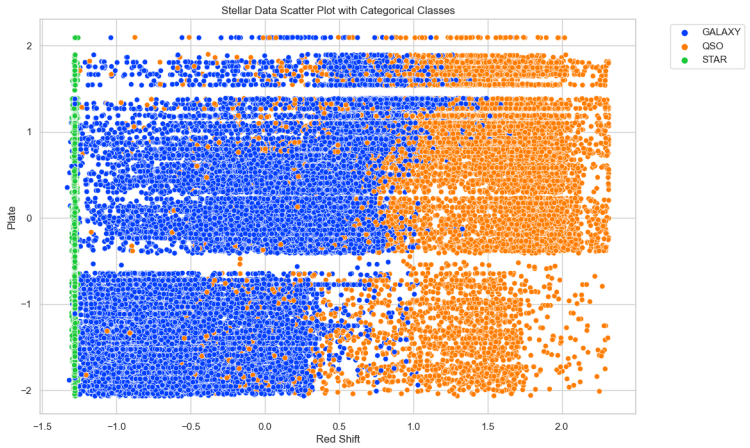

Figure 2

Stellar Data Scatter Plot with Categorical Classes (X-axis: red shift, Y-axis: plate)

Supervised Machine Learning Models

K-Nearest Neighbors (KNN) Classification Model

The K-Nearest Neighbors (KNN) model, combined with feature scaling and grid search optimization, effectively classified celestial objects into galaxies, quasi-stellar objects (QSOs), and stars using Sloan Digital Sky Survey data. Feature normalization using StandardScaler was crucial for the distance-sensitive KNN algorithm. The optimal model, using 5 neighbors as determined through grid search with 5-fold cross-validation, achieved a high test accuracy of 96.07% and a similar cross-validation score of 95.95%. Overall, the model demonstrated robust performance with precision, recall, and F1-score all around 96%, confirming its effectiveness in classifying celestial objects.

Class	Precision	Recall	F1-Score	Support
GALAXY	0.96	0.97	0.97	5904
QSO	0.97	0.91	0.94	1887
STAR	0.95	0.98	0.96	2209
Accuracy			0.96	10000
Macro Avg	0.96	0.95	0.96	10000
Weighted Avg	0.96	0.96	0.96	10000

Figure 3

KNN Classification Report

Decision Tree Model

The second model evaluated in the report involves a Decision Tree Classifier, which is optimized using Grid Search to determine the best parameters for maximizing classification accuracy.

The best score from cross-validation is approximately 96.52%, indicating strong performance during the training phase. Meanwhile, the optimized decision tree model achieves an

accuracy of 96.65% on the test set, slightly higher than the cross-validation score, showcasing its effectiveness on unseen data.

The model demonstrates high precision and recall across all categories, with a macro average precision of 97% and recall of 95%.

The F1-Score, which balances precision and recall, is also notably high, averaging 96% across all classes, with stars achieving a perfect score of 1.00.

Class	Precision	Recall	F1-Score	Support
GALAXY	0.96	0.99	0.97	5904
QSO	0.96	0.86	0.91	1887
STAR	1.00	1.00	1.00	2209
Accuracy			0.97	10000
Macro Avg	0.97	0.95	0.96	10000
Weighted Avg	0.97	0.97	0.97	10000

Figure 4

Decision Tree Classification Report

Random Forest Model

The third model explored in the analysis utilizes a Random Forest Classifier, a powerful ensemble learning method known for its high accuracy and robustness in handling complex classification tasks. The model is first set up with a basic configuration and then fine-tuned using Randomized Search to optimize its parameters.

The best model from the Randomized Search achieves an impressive cross-validation score of 98%, suggesting consistent and reliable performance across different subsets of the training data. The fine-tuned model confirms its effectiveness with a test accuracy also at 98%, matching the cross-validation insights.

Class	Precision	Recall	F1-Score	Support
GALAXY	0.98	0.99	0.98	5904
QSO	0.97	0.93	0.95	1887
STAR	0.99	1.00	1.00	2209
Accuracy			0.98	10000
Macro Avg	0.98	0.97	0.98	10000
Weighted Avg	0.98	0.98	0.98	10000

Figure 5

Random Forest Classification Report

AdaBoost Model

Model 4 utilizes the AdaBoost (Adaptive Boosting) algorithm combined with Decision Trees to classify celestial objects into three categories: galaxies, quasars (QSO), and stars. AdaBoost is an ensemble technique that combines multiple weak learners (in this case, decision trees) to create a strong classifier. By focusing on the mistakes of previous learners, AdaBoost aims to improve the model's overall predictive accuracy.

The best AdaBoost model achieved a cross-validation score of 93% and replicated the initial test set accuracy of 80.93%.

The overall accuracy of AdaBoost Model stood at 98%, demonstrating high effectiveness across all classes.

Macro and weighted averages for precision, recall, and F1-score were all impressively high, around 98%.

Class	Precision	Recall	F1-Score	Support
GALAXY	0.98	0.99	0.98	5904
QSO	0.97	0.93	0.95	1887
STAR	0.99	1.00	1.00	2209
Accuracy			0.98	10000
Macro Avg	0.98	0.97	0.98	10000
Weighted Avg	0.98	0.98	0.98	10000

Figure 6

AdaBoost Classification Report

Unsupervised Machine Learning Models

K-Means Model

Based on the elbow method findings, a K-Means model was configured with three clusters. To illustrate the result, scatter plots are generated. The first plot demonstrated relatively clear cluster separation that depicts the clustering based on the right ascension angle versus near infrared filter, suggesting these features are effective for distinguishing between groups.

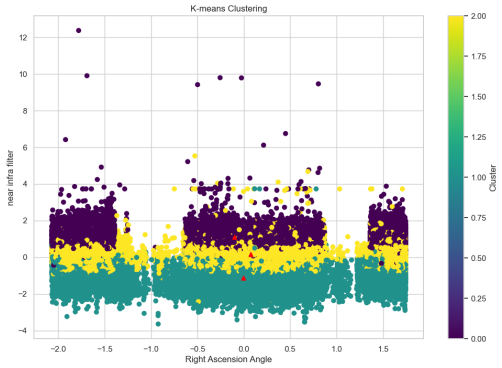

Figure 7

Scatter plot depicts the clustering based on the right ascension angle versus near infrared filter

The second plot also shows clusters based on right ascension angle and declination angle. However, the clusters in this plot appeared more blended, suggesting that the data's separability based solely on these features might be limited.

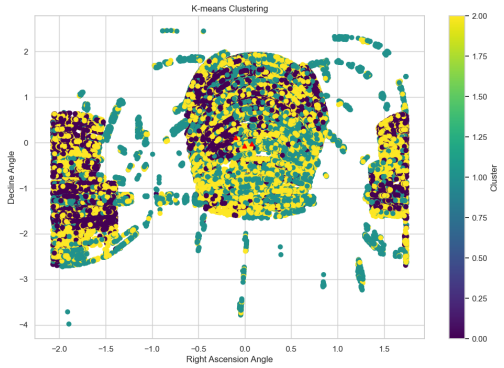

Figure 8
Cluster plot based on right ascension angle and declination angle

As for the third group of pairwise plots, we notice that most pairwise plots reveal clusters that are not well distinguished, with some being very close to each other, which could indicate overlapping properties among the objects or insufficient feature discrimination for the chosen clustering approach.

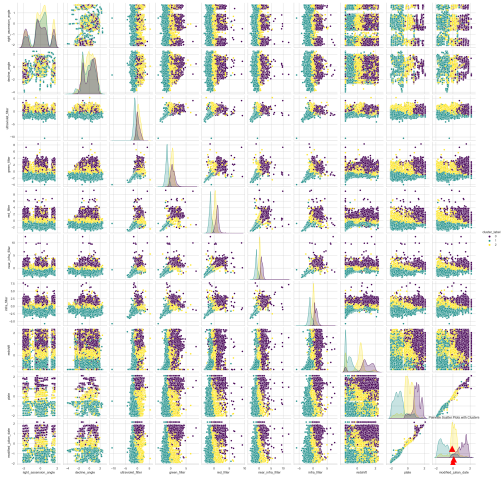

Figure 9
Pairwise plots

For the performance score, given as a negative sum of squared distances from points to their respective cluster centroids (SSE), was -487102.98. This score reflects the spread of the clusters, with a lower (or more negative) value typically indicating fewer compact clusters.

Although the K-Means clustering provided useful insights into the astronomical dataset, particularly highlighting how certain features might be more effective in grouping celestial objects, the primary challenge in applying K-Means clustering to this dataset is the complexity and potential non-linearity of astronomical data distributions. The clusters, especially evident in figures where feature differentiation was less apparent, were not well-separated.

Findings

To sum, the Random Forest and AdaBoost models showed outstanding precision and recall, especially for the star class, which achieved perfect scores.

Decision Tree and KNN models demonstrated slightly lower performance metrics but remained highly effective, particularly in accurately classifying galaxies and stars.

Consistently, Random Forest and AdaBoost models exhibited high F1-scores, reflecting a balanced performance between precision and recall.

Decision Tree and KNN maintained competitive F1-scores across all classes, especially for galaxies and stars.

Model	Class	Precision	Recall	F1-Score	Model	Class	Precision	Recall	F1-Score
KNN Model	GALAXY	0.96	0.97	0.97	**Random Forest**	GALAXY	0.98	0.99	0.98
	QSO	0.97	0.91	0.94		QSO	0.97	0.93	0.95
	STAR	0.95	0.98	0.96		STAR	0.99	1.00	1.00
	Accuracy			0.96		Accuracy			0.98
	Macro Avg	0.96	0.95	0.96		Macro Avg	0.98	0.97	0.98
	Weighted Avg	0.96	0.96	0.96		Weighted Avg	0.98	0.98	0.98
Decision Tree	GALAXY	0.96	0.99	0.97	**AdaBoost**	GALAXY	0.98	0.99	0.98
	QSO	0.96	0.86	0.91		QSO	0.97	0.93	0.95
	STAR	1.00	1.00	1.00		STAR	0.99	1.00	1.00
	Accuracy			0.97		Accuracy			0.98
	Macro Avg	0.97	0.95	0.96		Macro Avg	0.98	0.97	0.98
	Weighted Avg	0.97	0.97	0.97		Weighted Avg	0.98	0.98	0.98

Figure 10

Comparison of different models' precision, recall and F1 scores

Considering the test accuracy, precision, recall, and F1-scores, the Random Forest model is identified as the best model due to its highest accuracy (98.00%) and excellent balance between precision and recall across all classes. It also exhibited the highest macro and weighted averages for the F1-score.

Meanwhile, the Decision Tree model serves as a robust alternative with nearly comparable performance metrics, slightly lower complexity, and excellent interpretability.

For the AdaBoost Model, although it showed perfect metrics, its overall accuracy significantly lagged behind the other models, potentially due to overfitting or the influence of noisy data on the boosting process.

The K-Means Clustering model, while not directly comparable due to its unsupervised nature, provided valuable insights into the natural groupings within the data, useful for exploratory data analysis and preliminary clustering before applying supervised models.

In conclusion, the Random Forest model stands out as the most effective model for classifying astronomical objects based on the Sloan Digital Sky Survey data. It combines high accuracy with robustness against overfitting and provides detailed insights into the importance of features used for classification. Future work could explore deeper into hybrid models or more complex ensemble techniques to potentially enhance prediction accuracy further.

Note: All models in this report are run in MacBook, the results might be different in Windows.