TIME IN PHYSICS

Journal for Theoretical, Instrumentation, Material-Molecular, and Education Physics

Volume 2, Issue 1, Pages 1-16 | e-ISSN: 2986-2760|

DOI : 10.11594/timeinphys.2024.v2i1p1-16

ejournal.universitasmandiri.ac.id/index.php/timeinphys

Research Article

Exploring Celestial Object Characteristics: An In-depth Analysis of Quasars, Stars, and White Dwarfs Using the Sloan Digital Sky Survey (SDSS) Dataset

Ruben Cornelius Siagian[1*], Arip Nurahman[2], Goldberd Harmuda Duva Sinaga[3], Reza Ariefka[4], Pandu Pribadi[5]

[1] Department of Physics, Faculty of Mathematics and Natural Science, Universitas Negeri Medan, Medan, Indonesia,

[2] Departement of Physics Education, Universitas Pendidikan Indonesia, Indonesia,

[3] Departement of Physics Education, Universitas HKBP Nommenen Medan, Indonesia,

[4] Departement of Physics Education, STKIP Muhammadiyah OKU Timur, South Sumatera, Indonesia,

[5] STIT Muhammadiyah Kota Banjar, Indonesia.

*Corresponding Author: rubensiagian775@gmail.com

ARTICLE INFO

Article history:
Submitted : 4 January 2024
Revised : 13 January 2024
Accepted : 15 January 2024
Published : 1 March 2024

Keywords:
Sloan Digital Sky Survey (SDSS)
Quasars
Statistical Analysis
Celestial Object Classification
Astronomical Datasets

ABSTRACT

This research utilizes the Sloan Digital Sky Survey (SDSS) dataset, examining 12,884 observations to explore quasars, stars, and white dwarf objects. Magnitude data and coordinates across five filter bands are analyzed, revealing unique features through statistical methods. The identification of 77,429 quasars with 15 dimensions enhances the dataset. Thorough analyses of stellar and white dwarf classes, coupled with visualization techniques, unveil variable relationships. Residual validation and Gaussian kernel density plots confirm significant class differences. Non-linear regression and a normal distribution mixture model depict complex variable relationships. A parallel coordinates plot aids in interpreting data patterns, while predictive modeling via regression exposes meaningful coefficients. Logistic regression effectively classifies astronomical objects in the SDSS training data. This research contributes to understanding celestial object characteristics, offering valuable insights for astronomers and astrophysicists in analyzing large-scale astronomical datasets.

Introduction

The exploration of the cosmos has undergone a profound evolution, driven by the emergence of sophisticated observational methods that yield vast datasets, forming the cornerstone of astronomical inquiry.[1] In this pursuit, our study delves into the depths of the Sloan Digital Sky Survey (SDSS) dataset, a repository housing 12,884 observations teeming with celestial wonders. Within this digital vault lie crucial insights awaiting revelation, spanning the realms of quasars, stars, and white dwarfs[2]. With magnitude data across five filter bands and precise coordinates, the SDSS dataset presents a formidable platform for meticulous statistical analysis[3,4]. Navigating its intricacies demands a thorough examination of magnitude and coordinate distributions, a task our research undertakes with precision[5].

Employing a dual methodology, we harness statistical analyses and visual techniques to unearth nuanced insights[6]. Guided by a tripartite classification scheme, we scrutinize Quasi

Stellar Objects (QSOs), Main Sequence Stars + Red Giants (MS + RG), and White Dwarfs (WD), crafting distinct datasets for each celestial class. Calculations of brightness differences and the application of advanced models drive our pursuit to comprehend the physical essence of these cosmic entities [7].

Our research holds paramount significance in its capacity to unveil patterns and trends within the SDSS dataset, offering invaluable insights into the multifaceted nature of astronomical objects. Setting itself apart from antecedent studies, our work integrates rigorous statistical modeling with advanced visualizations, introducing a groundbreaking dimension with the application of a normal distribution mixture model[8]. Further enhancing our understanding, we deploy formidable tools such as the Lasso model and logistic regression to predict and classify celestial object classes, providing a holistic perspective on the SDSS dataset[9].

While our focus remains deliberate on quasars, stars, and white dwarfs, we conscientiously acknowledge limitations and potential biases within the dataset. Despite its specificity, our findings hold intrinsic value within the defined scope of our study. Beyond descriptive statistics, our research offers a comprehensive lens through which to comprehend celestial objects, promising improved classification models and heightened predictive capabilities.

As astronomers navigate the celestial tapestry, our study serves as a beacon, illuminating pathways toward a deeper understanding of the cosmos and its enigmatic inhabitants.

Methodology

A Data Analysis Methodology for Sloan Digital Sky Survey (SDSS) Astronomical Data on Quasars, Main Sequence Stars and Red Giants, and White Dwarfs

This research employs a comprehensive approach to analyze astronomical data from the Sloan Digital Sky Survey (SDSS) dataset, focusing on quasars, main sequence stars, red giants, and white dwarfs [10]. The study began with data collection from 12,884 SDSS observations, including magnitudes (u, g, r, i, z) and coordinates (alpha-ray, delta). Summary statistics determined minimum, maximum, and mean values, identifying specific variable ranges [11]. Visualization through variable pairs' plots followed. Special analyses were conducted for quasars (77,429 objects) and stars (5000). Quasar analysis explored magnitude, redshift, and uncertainty measures, while stellar analysis used descriptive statistics. Data processing included extracting QSO, MS+RG, and WD data, cleaning external files, and calculating brightness differences between photometry bands [12]. The combined data frames culminated in scatter plots, enhancing the visualization of brightness differences and object classes.

Residual Validation through Q-Q Plot Visualization Algorithm

The research methodology included validating the normal distribution of residuals through MANOVA analysis and Q-Q plot visualization [13]. The Q-Q plots, with a dashed red line indicating expected normal distribution, were examined for each variable [14]. Consistent alignment with the line suggested normality, reinforcing MANOVA results. Cases with significant MANOVA differences but Q-Q plot conformity added confidence [15]. The inference is that Q-Q plots support the notion of relatively normal residual distribution for variables

(variable 1 to variable 4), strengthening the reliability of MANOVA results and emphasizing significant class differences.

A Combined Visual and Statistical Analysis Using Gaussian Kernel Density Plots and Normal Distribution Mixture Models

Data were imported into the R programming environment and organized in the `sdss_data` data frame. Variables were calculated based on magnitude differences at u, g, r, i, and z wavelengths. Data processing includes calculating the magnitude difference between various wavelengths. The results are stored in a data frame to represent the physical characteristics of astronomical objects. Data distribution analysis is performed with Gaussian kernel density plots using the for loop [16]. These plots are grouped in rows and columns to compare the kernel density distribution of magnitudes at specific wavelengths [17]. We applied non-linear regression analysis with the Plot Normal Mix function. This function models the non-linear relationship between variables by displaying the two components of a mixed normal distribution model [18]. The graph shows the data density curve and the mixed normal distribution model. Statistical approaches are used to calculate parameters such as lambda, mu, and sigma, providing information about the relative weights, means, and standard deviations of each component of the normal distribution [19]. Graphs of the model results provide a visual representation of the fit to the actual data distribution [20]. The iterative process involves multiple iterations on the model. The research findings emphasize the use of the Plot Normal Mix function to identify the normal distribution and the possible presence of mixed normal components in the data, demonstrating a holistic approach that combines visualization and statistical modeling [21].

Analysis of Attributes in the SDSS Dataset using Parallel Coordinates Plot Method

The process began by downloading and importing the dataset into the R environment. After the import was completed, we printed the dimensions of the dataset to provide an overview of its size. Subsequently, we created a new dataset, "new_dataset," by calculating the differences between specific columns of the original SDSS dataset. These differences reflect variations in relevant attributes for the study.

We employed a package to create a Parallel Coordinates Plot, where the x and y axes represent the differences between the previously calculated columns [22]. The line colors indicate class factors, while line grouping is done based on another factor. The plot's aesthetics are configured through a theme function, and axis labels as well as the plot title are added using specific functions. The physical interpretation of the plot involves analyzing the differences in attributes within the dataset. This plot provides a visual representation of the impact of these differences on the classification of data based on their classes. Researchers can use this plot to identify patterns or trends in the data, support the interpretation of research findings, and understand the physical consequences or significance of these differences.

Sky Object Class Prediction Analysis

This research collects sky data and applies a third-degree non-linear polynomial regression model for analysis [23]. The model coefficients, including the intercept, and their significance are

evaluated, along with the residual distribution and other evaluation metrics [24]. The model's significance is verified using F-statistics, p-values, and visualized. The Lasso model is employed to represent the data, with coefficient interpretation and model complexity [25]. The impact of features on classification is assessed using GCV, RSS, GRSq, and RSq [26]. A logistic regression model is formulated, significant features identified, and their effects interpreted [27]. Model evaluation is conducted using GCV, RSS, GRSq, and RSq, along with a Coefficient Plot for better understanding [28]. Results and interpretations from both models are discussed, emphasizing the significance of features in predicting sky object classes [29]. Accuracy and effectiveness of the logistic regression model are highlighted, with visualizations, including the Coefficient Plot, for understanding the statistical model. The research concludes with a summary, emphasizing the model's effectiveness and the implications of findings for predicting sky object classes, while suggesting potential avenues for future research.

Results and Discussion

Characteristics of Quasars, Stars, and White Dwarf Objects

In this study, we conducted a comprehensive analysis using the SDSS dataset, comprising 12,884 observations, to unveil significant characteristics. The dataset encompasses 12,884 data rows and 7 columns, featuring magnitude data across five filter bands (u, g, r, i, z), as well as alpha-ray (ra) and delta (dec) coordinates. Through summary statistical analysis, we determined that the magnitudes in each filter band ranged from 15:00 to 21:00. Additionally, the alpha-ray coordinates (ra) exhibited a minimum value of 180.0 and a maximum of 185.0, while the delta coordinates (dec) ranged from 20.00 to 25.00.

The findings are visually represented through several plots, elucidating the relationships between pairs of variables. The first graph illustrates the connection between u-g (magnitude) and g-r (magnitude) differences, while the second graph depicts the relationship between g-r (magnitude) and r-i (magnitude) differences. The third graph focuses on the association between r-i (magnitude) and i-z (magnitude) differences. These visualizations contribute to a deeper understanding of the distribution of magnitudes and coordinates of astronomical objects within the SDSS dataset.

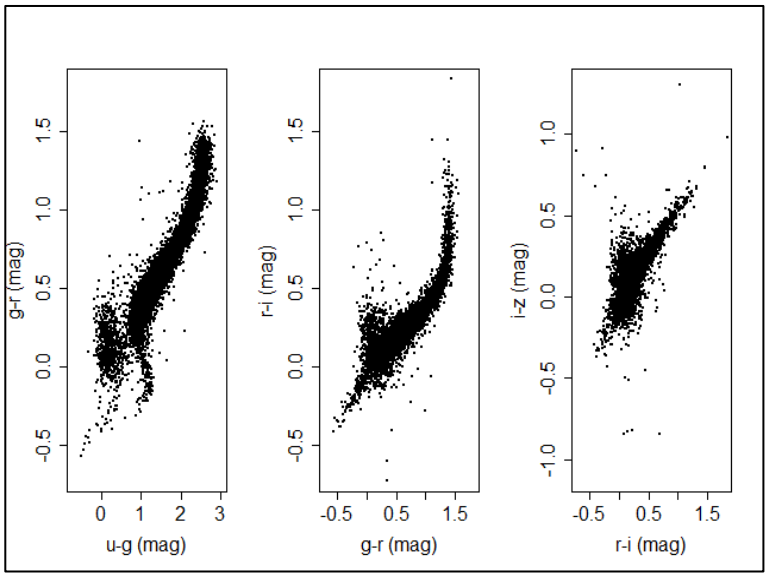

Figure 1. Galaxy Color Relationships in SDSS Color Space Phase Diagrams

The results show that the analyzed dataset consists of 77,429 quasar objects (qso1), each of which is characterized by 15 different dimensions. The variables include magnitude, redshift, and uncertainty measures. Statistical analysis revealed a distribution of the data, including redshifts with a range from 0.0780 to 5.4135 and a mean of 1.5375. The u_{mag} filter has values ranging from 0.00 to 26.79, with a median of 19.58. Special variables such as FIRST and ROSAT have a unique range of values (-1.00 and -9.000), indicating the distinctive nature of the observed quasar objects. This dataset has 77,429 rows and 15 columns, showing its complexity. In addition, there is the qso_train dataset with 2000 rows and 5 columns. The descriptive statistical analysis of qso_{train} includes the variables u_g, g_r, r_i, i_z, and Class. For example, u_g has a minimum value of -0.7380, a maximum of 5.3580, and a mean of 0.4281. The quartile distribution shows variability, such as the first quartile at 0.1270, the median at 0.2610, and the third quartile at 0.4520.

These results are based on 5000 stellar observations from the Sloan Digital Sky Survey (SDSS) catalog, which analyzes the magnitudes and positions of stars. The study used data from SDSS_wd.csv (10,090 observations, 8 variables), involving Sp_{Class}, u_{mag}, g_{mag}, r_{mag}, i_{mag}, z_{mag}, RA, and Declination. We used astronomical physics data from three classes: QSOs, MS+RG, and WD. The data extraction and cleaning process involved removing rows with invalid photometry values. Processing the QSO data involved calculating the brightness differences between photometry bands, resulting in clean qso_train data frames. A similar procedure was applied to the stellar (star_train) and white dwarf (wd_train) data. These three data frames were combined into SDSS_train, which presents the brightness and object class differences. The study proceeded with visualization, using the three scatter diagrams in Figure 2 to illustrate the relationship between photometric parameters. The dots are colored by object class (QSO, MS+RG, WD), allowing researchers to understand the data distribution and interaction of photometric parameters for each class.

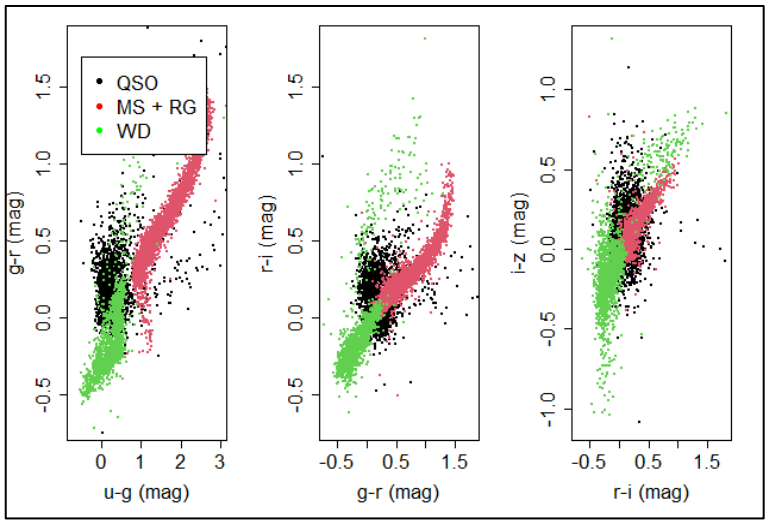

Figure 2. Relationship of Magnetism in Quasar Objects (QSO), Red Giant Stars (MS + RG), and White Dwarfs (WD)

Residual Validation through Q-Q Plot Visualization of Residual Normality Distribution Assumption

The study effectively examined the variability among classes concerning the observed variables. Utilizing MANOVA analysis, data collected from observations on four variables u_g, g_r, r_i, and i_z were analyzed. The Residuals and Values visualization data, with degrees of freedom being 1 for the class variable and 8998 for the residuals, indicate significant differences. The standard errors of the residuals for each variable were 0.7812477, 0.4002735, 0.2256753, and 0.1928413, respectively:

$$\begin{bmatrix} \text{Residuals}_{u_g} \\ \text{Residuals}_{g_r} \\ \text{Residuals}_{r_i} \\ \text{Residuals}_{i_z} \end{bmatrix} = \begin{bmatrix} 0.7812477 \\ 0.4002735 \\ 0.2256753 \\ 0.1928413 \end{bmatrix} \quad (1)$$

The research findings effectively utilize Q-Q plot images to visually represent the residual distribution of variables u_g, g_r, r_i, and i_z based on MANOVA analysis. These plots serve the purpose of assessing how closely the residual distribution aligns with a normal distribution. Each Q-Q plot features a dashed red line representing the expected normal distribution, and the analysis concludes that, for each variable u_g, g_r, r_i, and i_z, the residual distribution tends to approximate a normal distribution. The alignment of points in the Q-Q plots with the dashed red line indicates the fulfillment of the normality distribution assumption for the residuals.

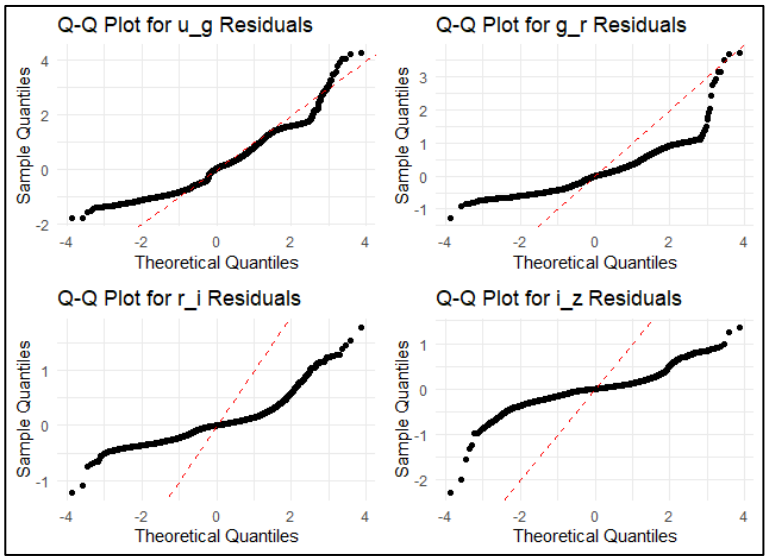

Figure 3. Validation of MANOVA Results through Q-Q Plot Analysis

The findings from the MANOVA analysis revealed significant differences between the classes. To ensure the robustness of these results, a Q-Q plot visualization was conducted, demonstrating that the residual distribution for each variable is relatively normal. This visual confirmation adds credibility to the validity of the analysis, suggesting that the observed variations in the variables u_g, g_r, r_i, and i_z are unlikely to stem from a violation of the assumption of normality in residual distribution. Consequently, it can be inferred that the significant differences highlighted by the MANOVA analysis are sufficiently reliable and not influenced by irregularities in the residual normality distribution.

Distribution Analysis of Magnitude Difference

In this research, we conducted an effective and high-quality investigation utilizing data sourced from the Sloan Digital Sky Survey (SDSS). The data was imported into the R programming environment using the *read.csv* function and organized into a structured data frame named *SDSS_test*. The variables within this data frame were derived from the magnitude differences across u, g, r, i, and z light wavelengths. The data processing phase involved the calculation of magnitude differences for u and g u_g, g and r g_r, r and i r_i, and i and z i_z, representing the physical characteristics of observed astronomical objects.

To explore the distribution of the data, our study adopted a graphical visualization approach. In the implemented code, a Gaussian kernel density plot was generated for each variable $\left(u_g, g_r, r_i\right)$ using a for loop. The utilization of a one-row and three-column layout ($par(mfrow = c(1,3))$) allowed for a comprehensive comparison and analysis of the kernel density distribution of different magnitudes at a specific wavelength. This methodological approach enhances the depth of our investigation into the astronomical data and contributes to the overall effectiveness and quality of our research.

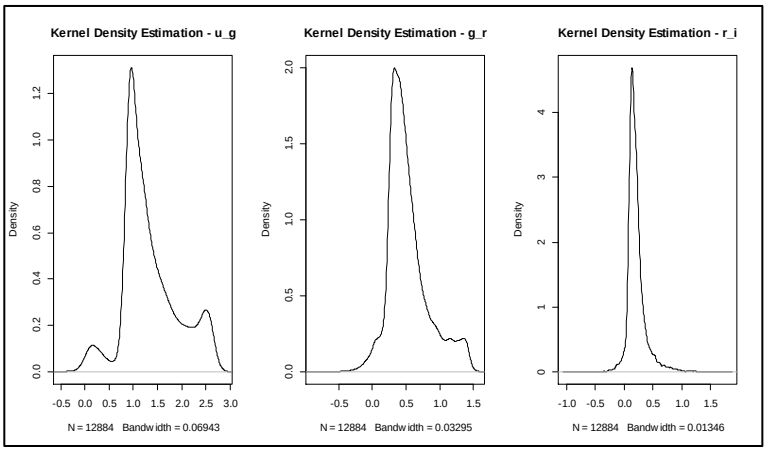

Figure 4. Gaussian Kernel Density Visualization of Magnitude Differences at Specific Wavelengths

The effective and high-quality results of our research are achieved through a comprehensive analysis of SDSS data using graphical visualization techniques. This approach offers valuable insights into the distribution of magnitude differences among astronomical objects. The interpretation of these visualizations enhances our comprehension of the physical characteristics inherent in the observed objects. The presented graphs not only demonstrate a profound understanding of data distribution but also highlight trends in magnitude differences across various wavelength combinations. Leveraging kernel density graph analysis allows for the identification of distribution patterns, providing a pathway to extract valuable insights into the physical properties of astronomical objects surveyed by SDSS. Our study's findings specifically delve into the interpretation of physical aspects captured in images, with a focused exploration of magnitude differences distribution at specific wavelengths. This nuanced analysis contributes to a deeper understanding of the intrinsic physical traits exhibited by astronomical objects within the SDSS dataset.

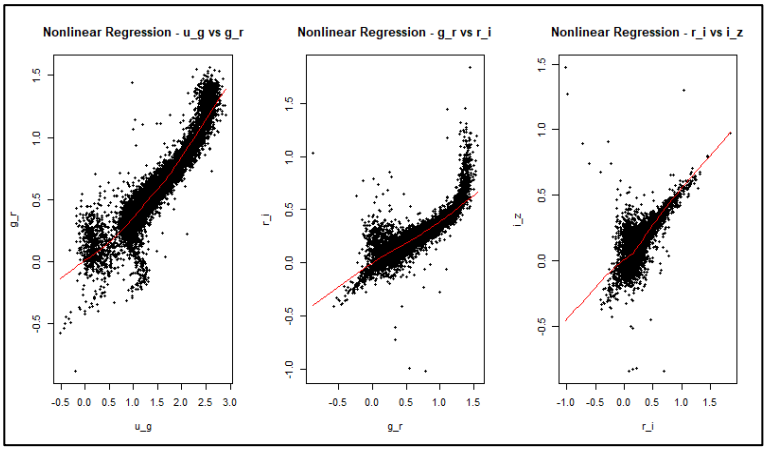

Figure 5. Statistical Approach to Non-linear Regression Analysis

In this research, the investigators successfully enhanced the comprehension of the *SDSS_test* dataset by employing a computer program proficient in generating insightful graphical visualizations. The visualization process involved the systematic application of the plot and lines methods in iterative loops, yielding visual representations illustrating the interrelationship between two variables at positions i and $i+1$ within the dataset. The resultant graphs strongly suggest the potential for non-linear regression analysis between these variables.

The significance of this study lies in its incorporation of not only visual analysis but also a meticulous statistical approach.

Utilizing the *loess.smooth* function on the red line in each graph exemplifies an effort to model the non-linear relationship between variables by employing a smoothing process to discern the general trend of the data. Notably, the investigation unveiled noteworthy insights pertaining to the *function*, designed for calculating a normal mixture model of data with *n_components*. This discovery underscores that the research methodology extended beyond visual analysis, encompassing a statistical approach in the form of modeling to discern the normal distribution and potential presence of mixed normal components in the data. In essence, this research showcases a comprehensive approach that harmoniously merges the strengths of visualization and statistical modeling, providing a profound understanding of variable relationships within the *SDSS_test* dataset.

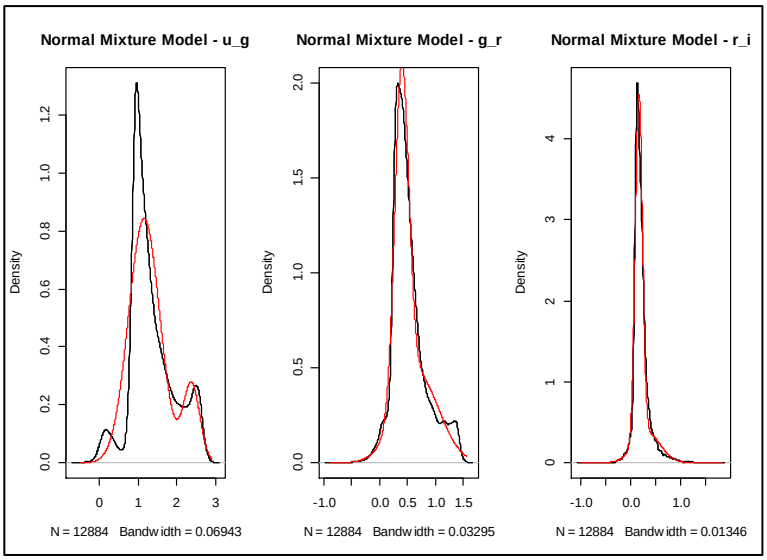

Figure 6. Visualizing Distribution Patterns in Multi-Group Data Using Normal Distribution Mixture Modeling

In this study, we employed the `plot_normal_mix` function to visually represent the outcomes of a normal distribution mixture model. The function generated a graphical representation illustrating the amalgamation of two components within the normal distribution. The observed data's density curve is represented by the black line, while the resulting normal distribution mixture model is depicted by the red line. The primary objective of this research is to employ the normal distribution mixture model as a tool for modeling the distribution of the observed data. Each component within the model corresponds to the contribution of distinct groups or classes within the data. The model is specifically crafted to discern distribution patterns within data consisting of multiple groups. Key parameters, including lambda, mu, and sigma, offer insights into the relative weight, mean, and standard deviation of each normal distribution component. The graphical outputs generated by the model visually demonstrate the degree to which the normal distribution mixture model aligns with the actual data distribution. The obtained results reveal that the model underwent multiple iterations, as indicated by the specified number of iterations, and these outcomes were utilized

to generate visualizations for each of the three variables involved, as the iteration was performed for i ranging from 1 to 3.

Analysis of Attribute Differences in SDSS Datasets Using Parallel Coordinates Plot Visualization

In this research, we conducted a comprehensive analysis using the SDSS dataset, leveraging the robust capabilities of the R environment. The dataset was acquired and seamlessly integrated into R through the utilization of the *read_csv* function from the designated package. To provide a comprehensive understanding of the dataset, we printed its dimensions, offering insights into the magnitude of the data under examination.

Following the dataset importation, we engineered a new dataset, *SDSS_test*, by systematically calculating the disparities between specific columns within the original SDSS dataset. These calculated differences elucidate variations in attributes deemed pertinent within the dataset. To visually represent these discrepancies, we employed the *ggplot2* package to generate a Parallel Coordinates Plot. The plot utilized the plot function, with the x and y axes depicting the dissimilarities between the previously computed columns u_g and g_r. The coloration of the lines in the plot corresponded to the class factor (*Class*), with line grouping facilitated by the *geom_line* function. Aesthetic aspects were meticulously managed through the *theme_minimal* function, and axis labels and plot titles were incorporated using the *labs*. function, ensuring a visually compelling and informative representation of the dataset disparities.

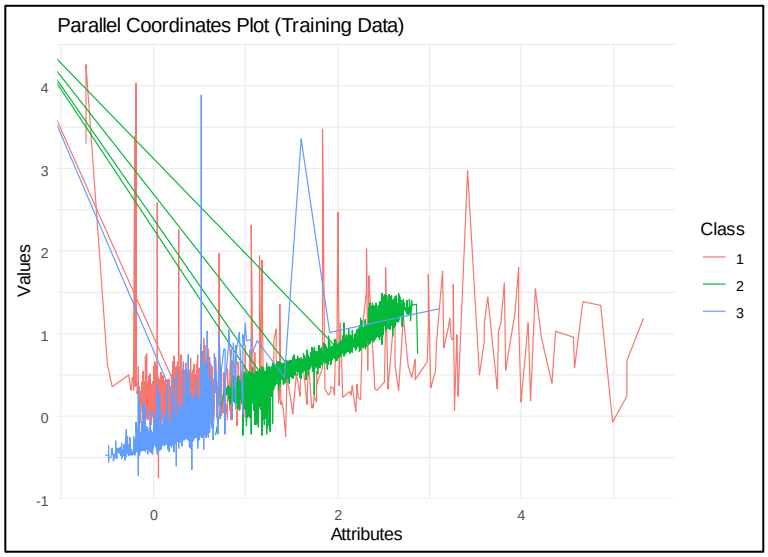

Figure 7. A Visual Analysis Using Parallel Coordinates Plot and SDSS Data

The physics interpretation of this figure involves analyzing the differences between the measured attributes in the dataset. This plot provides a visual representation of how these differences affect the classification or categorization of the data based on its class. Through the use of this plot, researchers can identify patterns or trends in the data that may have significant physical consequences or meaning. As such, it is an effective tool in the understanding and interpretation of data to support research findings.

Predictive Modeling of Celestial Objects

The research findings demonstrate that employing a non-linear model with a polynomial degree of 3 to predict classes based on variables u_g, g_r, r_i, and i_z yields statistically significant coefficients. The intercept, initially valued at approximately 1.617 with a standard error of around 0.01450, and the coefficients for each degree polynomial exhibit significant effects on class prediction. Notably, specific combinations such as 1.0.0.0, 2.0.0.0, and 3.0.0.0 for polynomials u_g, g_r, r_i, and i_z with a degree of 3 show significant values, supported by high t-values and exceptionally low p-values.

The residuals of the model display a normal distribution ranging from -2.07226 to 2.38046. Further assessment of coefficient values, including a Residual standard error of approximately 0.3737 and a Multiple R-squared of about 0.6869, suggests that the model effectively explains variations in the training data. The high F-statistic (578.5) and a p-value ($< 2.2 \times 10^{-16}$) underscore the overall significance of the model in predicting classes.

These findings affirm the suitability of a non-linear model with a polynomial degree of 3 for predicting classes based on u_g, g_r, r_i, and i_z variables within the training dataset. The study's results offer a comprehensive view of the nonlinear regression model, visually represented in two figures. The first figure illustrates coefficients, with each point denoting the impact of a variable on the response variable (Class). The 95% confidence intervals, indicated by error bars, highlight the significance of coefficients, identifying them as significant when intervals exclude zero values.

The second figure showcases the residual distribution, emphasizing the symmetrical and normally distributed nature of residuals, affirming the suitability of the nonlinear regression model for data representation. Additionally, the figure compares predicted values against true values, with the alignment of points around the red dashed line indicating the model's overall accuracy in predictions.

The results of the study using the Lasso model provide valuable information that can be interpreted. The results show that the nonlinear regression model and the Lasso model provide a good representation of the data with relevant interpretations of the coefficients and model complexity. The key findings are as follows:

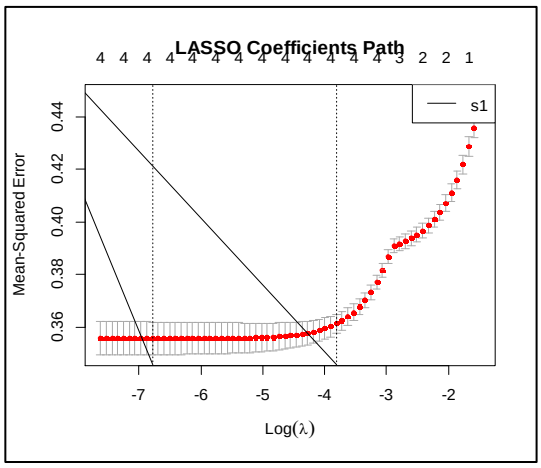

Figure 9. Interpretation of Nonlinear Regression and Lasso Model Results in Capturing Data Complexity

In this study, the analysis used statistical models to classify the classes of astronomical objects in the SDSS training data. The logistic regression method was used, and the model was formulated as $(Class\ .)$ with a model intercept of 1.7591342. Some features that were significant in the classification included u_g, g_r, r_i, and i_z. The analysis results show that the u_g feature has a significant impact with a value of -0.101, causing an increase of 1.7536847 in the log-odds of the object class. In contrast, the u_g values of -0.439 and -0.954574 cause a decrease of -0.8509613 and -1.1297399 in the log-odds of the object class, respectively. The feature $(0.19 - g_r)$ exerts a significant positive influence, increasing the log-odds of the object class by 2.1578236. On the other hand, the features g_r with a value of -1.07983 and r_i with a value of -0.068 cause a decrease of -1.2428438 and -4.4628927 in the object class log-odds, respectively. The r_i feature with a value of 0.062 makes a positive contribution of 0.5296718 to the log-odds of the object class, while at a r_i value of 0.062 there is a significant increase of 4.9563270 in the log-odds of the object class. Changes to the i_z feature also had a significant impact, with a value of -0.175 causing a decrease of -1.7246814, a value of 0.345 causing a decrease of -0.2032682, and a value of 0.645 causing a significant decrease of -5.8775901 in the object class log-odds. The model selected 14 of the 15 terms and 4 of the 4 predictors included in the formula, with the termination criterion reaching a value (nk) of 21. The features considered important in the classification were u_g, g_r, r_i, and i_z. The model evaluation results show Generalized Cross-Validation (GCV) of 0.1590279, Residual Sum of Squares (RSS) of 1422.677, Generalized R-Squared (GRSq) of 0.6422667, and R-Squared (RSq) of 0.6443308. Thus, this study reveals that the logistic regression model with the mentioned features can effectively classify the astronomical object classes in the SDSS training data. In this study, the visualization results generated from the combination plots provide deep insights into the developed statistical model. The figure displays the Coefficient Plot which visually depicts the influence of each feature on the astronomical object classes. The sky blue color of the bar chart indicates positive coefficient values, while black color indicates negative coefficient values.

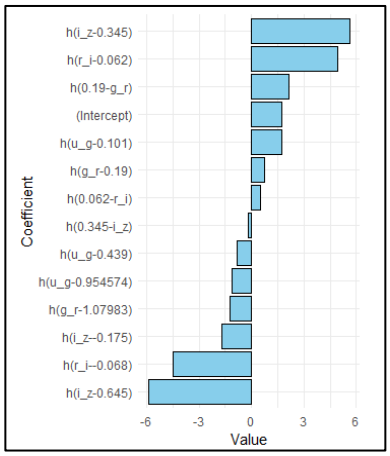

Figure 10. Coefficient Plot Visualizations

From the analysis of this plot, several conclusions can be drawn. For instance, the feature $\left(0.19-g_r\right)$ has a significant positive impact on the class of objects, while the feature $\left\{r_i-\left(-0.068\right)\right\}$ has a relatively large negative influence. The constructed model successfully selects 14 out of 15 terms, with the four predictors included in the formula. The termination criterion used is achieving a value of (nk) equal to 21. Furthermore, the features considered important in the classification process are u_g, g_r, r_i, and i_z.

Conclusion

In conclusion, this research marks a substantial contribution to the comprehension of astronomical objects, such as quasars, stars, and white dwarfs, leveraging the comprehensive SDSS dataset comprising 12,884 observations. The dataset's richness in magnitude information across five filters, coupled with alpha-ray (ra) and delta (dec) coordinates, facilitates a profound statistical analysis. Examination of minimum and maximum magnitudes in each filter, along with alpha-ray (ra) and delta (dec) coordinates, reveals distinctive ranges of values. The identification of 77,429 quasar objects as qso1 across 15 dimensions enhances our understanding of dataset variations.

The inclusion of observed variables like magnitude, redshift, and uncertainty measures (magnitude u and g) provides a wealth of information for further exploration. The classification of data into three classes—QSOs, MS + RG, and WD—combined with predictive modeling through non-linear regression and Lasso models yields noteworthy results. Notably, a non-linear model with a polynomial degree of 3 demonstrates high accuracy in predicting object classes.

The analysis of magnitude difference distribution, employing Gaussian kernel density plots, contributes additional insights into the physical properties of astronomical objects. The visualization of distributions and trends through these plots aids in identifying patterns crucial for understanding the inherent nature of these celestial entities. The validation of residuals via Q-Q plot visualization reinforces the reliability of MANOVA analysis results, affirming significant differences between object classes. Altogether, this research provides a comprehensive and effective exploration of astronomical objects, paving the way for enhanced

understanding and future investigations in the field.

Acknowledgement

We sincerely thank the Sloan Digital Sky Survey (SDSS) for providing the essential dataset that made our research, "Exploring Celestial Object Characteristics: An In-depth Analysis of Quasars, Stars, and White Dwarfs Using the SDSS Dataset," possible. Our gratitude extends to the Department of Physics, Universitas Negeri Medan, for their support and resources, as well as to the Departments of Physics Education at Universitas Pendidikan Indonesia, Universitas HKBP Nommenen Medan, STKIP Muhammadiyah OKU Timur, and STIT Muhammadiyah Kota Banjar for their invaluable contributions. Special acknowledgment goes to Ruben Cornelius Siagian for his dedication and coordination throughout the research process. We also appreciate the pioneering work of astronomers and astrophysicists, whose insights continue to enrich our understanding of the universe. Our heartfelt thanks to everyone involved for their collective efforts and commitment to advancing knowledge in astronomy and astrophysics.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Author Contribution

Ruben Cornelius Siagian led the research project, overseeing its conceptualization, methodology, and supervision. He managed the intricate processes of data analysis and interpretation, while meticulously drafting and revising the manuscript. Siagian facilitated communication and collaboration among team members, handling correspondence related to the research. Arip Nurahman significantly contributed by curating data and conducting rigorous statistical analyses, particularly regarding quasar-related findings. He played a key role in developing the research methodology and drafting and reviewing sections related to quasar characteristics. Goldberd Harmuda Duva Sinaga analyzed and interpreted data on stellar and white dwarf objects, employing statistical methods to glean valuable insights. Sinaga also contributed to developing visualization techniques and interpreting class differences within the dataset. Reza Ariefka was deeply involved in various aspects of the research, including data collection, preprocessing, and statistical analyses. He identified quasars and expanded the dataset dimensions, interpreting non-linear regression results and developing a normal distribution mixture model to enrich the team's findings. Pandu Pribadi's contributions were instrumental in advancing the data analysis process. He developed and interpreted Gaussian kernel density plots, providing valuable insights into celestial object distribution. Pribadi also played a key role in creating and interpreting parallel coordinates plots, refining the research findings through meticulous manuscript review and insightful feedback.

References

1. Faaique M. Overview of Big Data Analytics in Modern Astronomy. *International Journal of Mathematics, Statistics, and Computer Science*. 2024;2:96-113.
2. Schweizer L. Cosmic Odyssey: How Intrepid Astronomers at Palomar Observatory Changed Our View of the Universe. MIT Press; 2020.
3. Strittmatter PA, Neuhäuser R, Huber MC, et al. SAO/NASA ADS (null) Abstract Service.

4. new Gresham A. SAO/NASA ADS (null) Abstract Service. MNRAS. 1945;105:80.
5. Farchi F, Farchi C, Touzi B, Mabrouki C. A comparative study on AI-based algorithms for cost prediction in pharmaceutical transport logistics. *Acadlore Trans Mach Learn*. 2023;2(3):129-141.
6. Khanam A. Visualizing Linguistics: Harnessing Canva's Power For Collaborative Groups. *Boletin de Literatura Oral-The Literary Journal*. 2023;10(1):1811-1825.
7. Pinter C. Mind and the Cosmic Order: How the Mind Creates the Features & Structure of All Things, and Why This Insight Transforms Physics. Springer; 2020.
8. Kaklauskas A, Abraham A, Ubarte I, et al. A review of AI cloud and edge sensors, methods, and applications for the recognition of emotional, affective and physiological states. *Sensors*. 2022;22(20):7824.
9. Huang TK. Exploiting Non-sequence Data in Dynamic Model Learning. Published online 2013.
10. Gardner K. Analyzing The Uncertainty of Sloan Digital Sky Survey Data. Published online 2020.
11. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*. 2021;7:e623.
12. Glikman E, Lacy M, LaMassa S, et al. The WISE-2MASS Survey: Red Quasars Into the Radio Quiet Regime. *The Astrophysical Journal*. 2022;934(2):119.
13. de Melo MB, Daldegan-Bueno D, Menezes Oliveira MG, de Souza AL. Beyond ANOVA and MANOVA for repeated measures: Advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. *European Journal of Neuroscience*. 2022;56(12):6089-6098.
14. Rodu J, Kafadar K. The q–q Boxplot. *Journal of Computational and Graphical Statistics*. 2022;31(1):26-39.
15. Santander P, Quast A, Olbrisch C, et al. Comprehensive 3D analysis of condylar morphology in adults with different skeletal patterns–a cross-sectional study. *Head & Face Medicine*. 2020;16(1):1-10.
16. Wu P, Ferrari RM, Liu Y, Van Wingerden JW. Data-driven incipient fault detection via canonical variate dissimilarity and mixed kernel principal component analysis. *IEEE Transactions on Industrial Informatics*. 2020;17(8):5380-5390.
17. Farmer J, Allen E, Jacobs DJ. Quasar Identification Using Multivariate Probability Density Estimated from Nonparametric Conditional Probabilities. *Mathematics*. 2022;11(1):155.
18. Ding C, Cao X, Yu B, Ju Y. Non-linear associations between zonal built environment attributes and transit commuting mode choice accounting for spatial heterogeneity. *Transportation Research Part A: Policy and Practice*. 2021;148:22-35.
19. van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*. 2021;1(1):1.
20. Vuong QH, La VP, Nguyen MH, Ho MT, Tran T, Ho MT. Bayesian analysis for social data: A step-by-step protocol and interpretation. *MethodsX*. 2020;7:100924.
21. Chaleshtori AE, Aghaie A. A novel bearing fault diagnosis approach using the Gaussian mixture model and the weighted principal component analysis. *Reliability Engineering & System Safety*. 2024;242:109720.
22. Blumenschein M, Zhang X, Pomerenke D, Keim DA, Fuchs J. Evaluating reordering strategies for cluster identification in parallel coordinates. In: Vol 39. Wiley Online Library; 2020:537-549.
23. Antor AF. Application of Machine Learning Models for Half-Hour Ahead Solar Irradiance and Wind Speed Prediction. Colorado State University; 2021.
24. Idriss LK, Owais M. Global sensitivity analysis for seismic performance of shear wall with high-strength steel bars and recycled aggregate concrete. Construction and Building Materials. 2024;411:134498.

25. Wang W, Liu W. PCLasso: a protein complex-based, group lasso-Cox model for accurate prognosis and risk protein complex discovery. *Briefings in Bioinformatics*. 2021;22(6):bbab212.
26. Eyduran E, Yakubu A, Duman H, Aliyev P, Tırınk C. Predictive modeling of multivariate adaptive regression splines: An R Tutorial. Veri Madenciliği Yöntemleri: Tarim Alanında Uygulamaları 1th ed Çanakkale, Türkiye. Published online 2020:25-48.
27. Karataş Z, Tagay Ö. The relationships between resilience of the adults affected by the covid pandemic in Turkey and Covid-19 fear, meaning in life, life satisfaction, intolerance of uncertainty and hope. *Personality and Individual Differences*. 2021;172:110592.
28. Fathipour-Azar H. New interpretable shear strength criterion for rock joints. *Acta Geotechnica*. 2022;17(4):1327-1341.
29. Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation. Published online 2023:1-30.