COGS 118B - Project Proposal

Names

- Tianxing Zhou
- Samar Marwah
- Shaoming Chen
- Tianshi Hu
- Yueshan Huang

Abstract

Classifying galaxies, stars, and quasars has always been a pivotal problem in astrophysics research. **The project aims to distinguish astronomical objects such as galaxies, stars, or quasars(QSO)**. To effectively classify these objects, a precise measurement of photometry is required. We will use the Sloan Digital Sky Survey Data Release 18 (SDSS-DR18), which contains photometry information from ultraviolet to near-infrared. Our objective is to develop a machine-learning framework that employs dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), to distill the high-dimensional data inherent to astronomical observations. Subsequently, we tested and applied the KMeans Clustering algorithm, Gaussian Mixture Models (GMMs), and Hierarchical Clustering to classify the observed celestial bodies into distinct clusters. The performance of our classification model will be evaluated based on the Silhouette Score and Bayes Information Criterion (BIC). Furthermore, we explored supervised machine learning algorithms and built a Decision Tree Classifier with an accuracy of 98.58% on the test set. With the advent of next-generation telescopes such as the James Webb Space Telescope, the field of astronomy is in critical need of robust, automated classification methods capable of handling large and complex datasets. By using machine-learning techniques, we created a model capable of categorization astronomical objects, facilitating deeper insights into the cosmos.

Background

Exploring the depths of the cosmos and deciphering its mysteries lies at the heart of astrophysics. Central to this quest is the Sloan Digital Sky Survey (SDSS), a treasure trove of celestial data capturing the essence of stars, galaxies, and enigmatic quasars. This expansive dataset serves as a beacon for astronomers, offering profound insights into the universe's tapestry. With 100,000 observations, this dataset offers a comprehensive view of the universe's

constituents. Spectral characteristics, such as right ascension, declination, and photometric filter measurements across multiple wavelengths, are essential for understanding the nature of these celestial bodies. Additionally, redshift values derived from observed wavelength shifts provide crucial insights into their distances and velocities. Previous research utilizing SDSS data has significantly advanced our knowledge of cosmology, stellar evolution, and galaxy formation [1]. Leveraging machine learning approaches on this dataset enables automated classification, contributing to ongoing efforts in understanding the cosmos.

Problem Statement

In the context of modern astronomy, the handling and analysis of large and high-dimensional datasets of celestial objects are constrained by inherent data interpretation and object classification complexities. Data obtained from sky surveys, like that encompassing a mixture of stars, galaxies, and quasities, generate properties for celestial objects across several spectral bands. A critical condition is the ever-increasing rate of this data collection that outpaces the development of both computational training constraints and the granulation of the named cluster classes. Additionally, analyzing high-dimensional spectral and spatial features creates an encumbrance for creating intelligible 2D or 3D photometric representation models that can capture the underlying object type. Consequently, we intend to solve the problem statement by clustering new classes of astronomical objects, refining existing classifications based on their features, and Visualize high-dimensional data in 2D or 3D to identify patterns or groupings that are not apparent in higher-dimensional space with the help of K-Means, Hierarchical Clustering, Gaussian Mixture Models, t-SNE and PCA (potential solution).

Quantifiable: The variables in the dataset we used are numeric, hence we can utilize the mathematical expression to compute the variables and put them into K-means, Hierarchical Clustering even including PCA matrix calculations.

Measurable: Silhouette score in K-Means and Bayes Information Criterion for the Gaussian Mixture Models. Range: -1 to 1, where 1 indicates perfect agreement between clusterings.

Replicable: The dataset is replicable as the stars are very abundant in the universe and it is plausible to take another galaxy and find the star data and reproduce the result.

Data

- Link and reference: https://www.kaggle.com/datasets/diraf0/sloan-digital-sky-survey-dr18/data
- Description: This dataset contains data on celestial objects from the Sloan Digital Sky Survey, which comprises 100,000 observations with 42 features.
- Observation:

    - Objid, Specobjid - Object Identifiers
    - Ra: right ascension is used to pinpoint locations of objects in the sky.
    - Dec: declination is another coordinate used to pinpoint locations of objects in the sky, it measures the angular distance of an object north or south of the celestial equator.
    - Redshift: a measure of how much the wavelength of the light from an object has been stretched due to the expansion of the universe, indicating the distance of velocity of celestial objects moving away from us.
    - u, g, r, i, and z (Photometric Magnitudes):
        - u (ultraviolet): Captures light in the ultraviolet spectrum.
        - g (green): Measures light in the green part of the visible spectrum.
        - r (red): Measures light in the red part of the visible spectrum.
        - i (infrared): Detects light in the near-infrared spectrum.
        - z (near-infrared): Captures light in the near-infrared spectrum, but at longer wavelengths than 'i'.
    - run (Run Number): specific observational run which data was collected
    - rerun (Rerun Number): specific observational run which data was recollected
    - camcol (Camera Column): the column of the camera detector array that captured the image.
    - field (Field Number): an identifier for a specific region of the sky observed during the survey.
- Critical variables: Photometric Magnitudes
    - Measures the brightness of celestial objects in specific wavelength bands and each band captures light from a different part of the electromagnetic spectrum.
    - The magnitudes are presented as decimal numbers representing a ratio of brightness to a base celestial object (E.g. Vega). These magnitudes are measured on a logarithmic scale, defined such that a difference of 5 magnitudes corresponds to a factor of 100 in brightness. In other words, a decrease in magnitude by 1 unit represents an increase in brightness (and thus energy) by a factor of approximately 2.512. We can also calculate the magnitude using this formula: $m = -2.5 \log_{10} \left(\frac{F}{F_0} \right)$
- Data cleaning:
    - We will remove unimportant features such as Object Identifiers, run and rerun numbers, and field numbers, etc. Additionally, we will conduct standard tests to check for outliers, missing values, and duplicate entries. If necessary, we will also attempt to normalize the magnitude values to ensure they are on a comparable scale.

Proposed Solution

Our solution involves the application of K-Means and Hierarchical Clustering algorithms for pattern recognition and classification. PCA will be employed to transform the high-dimensional data into a lower-dimensional subspace so that we may preserve variance as much as possible. K-Means clustering will serve to partition the dataset into distinct groups, based on feature similarity. Hierarchical clustering will complement this process by providing a dendrogram, offering insights into the hierarchical structure of the data and allowing for the identification of nested clusters. This dual-clustering approach aims to maximize the interpretability of the data's structure in reduced dimensions and reveal subtle classifications that may not be discernible in the original high-dimensional space. The efficacy of this solution will be gauged by the clarity of the clusters formed and their alignment with known astrophysical phenomena.

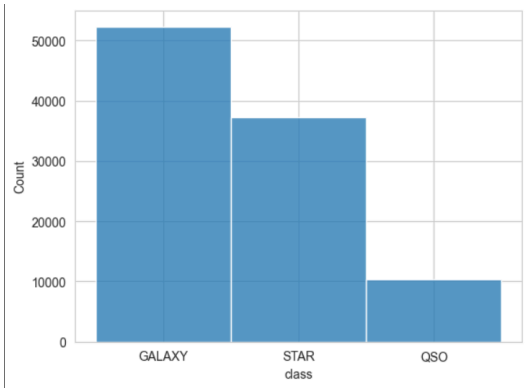
Evaluation Metrics

The evaluation metrics we used primarily are silhouette scores for the KMeans Clustering Algorithm and Bayes Information Criterion for the Gaussian Mixture Models. Silhouette score measures how similar an object is to its cluster compared to other clusters. Its range is -1 to 1, where 1 indicates that the sample is far away from neighboring clusters, suggesting that the sample is well placed within its cluster. Silhouette Score= $(b-a)/\max(a,b)$, where a is the mean distance between a sample and all other points in the same cluster, and b is the mean distance between a sample and all other points in the next nearest cluster (i.e., the cluster that is not its own but has the closest mean distance to the sample).

Results

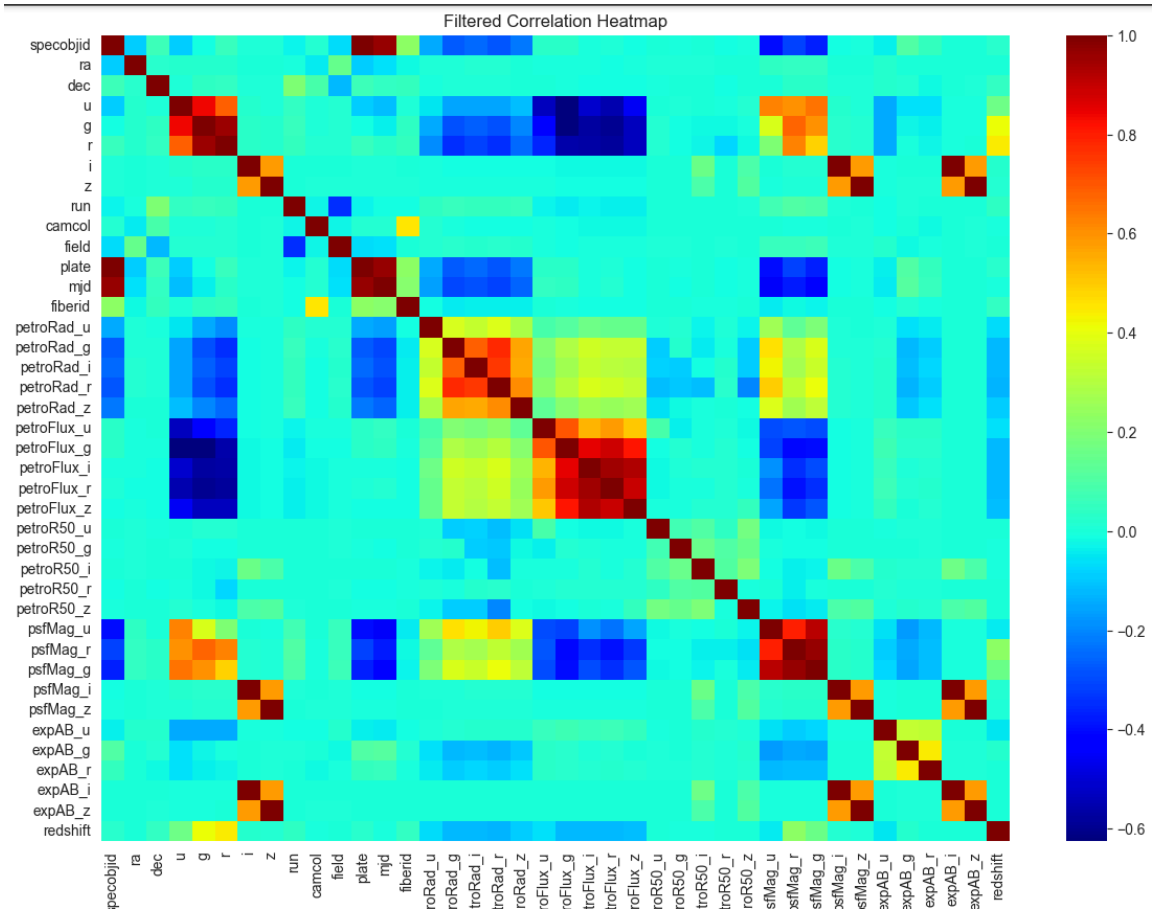
All relevant code attached to visualizations linked in the attached jupyter notebook

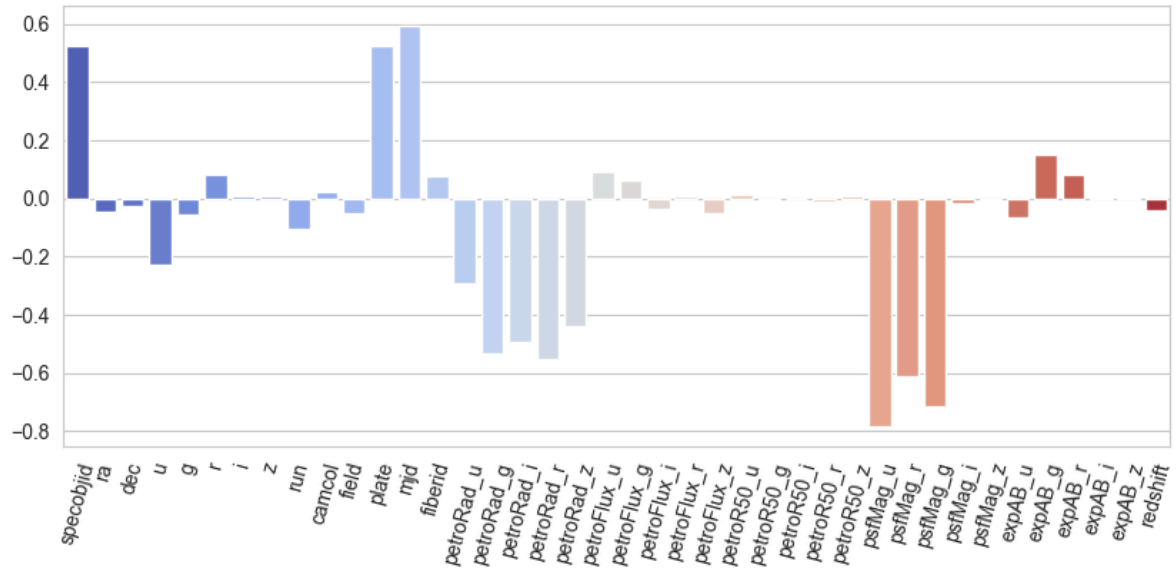
Our analysis focused on a dataset comprising galaxies, stars, and Quasi-Stellar Objects (QSOs). The dataset's composition was uneven, with galaxies being the most common class at approximately 50,000 counts, stars exceeding 30,000, and QSOs around 10,000. This imbalance necessitated careful consideration in our modeling approach to avoid biases towards the predominant class.



Correlation and Feature Redundancy

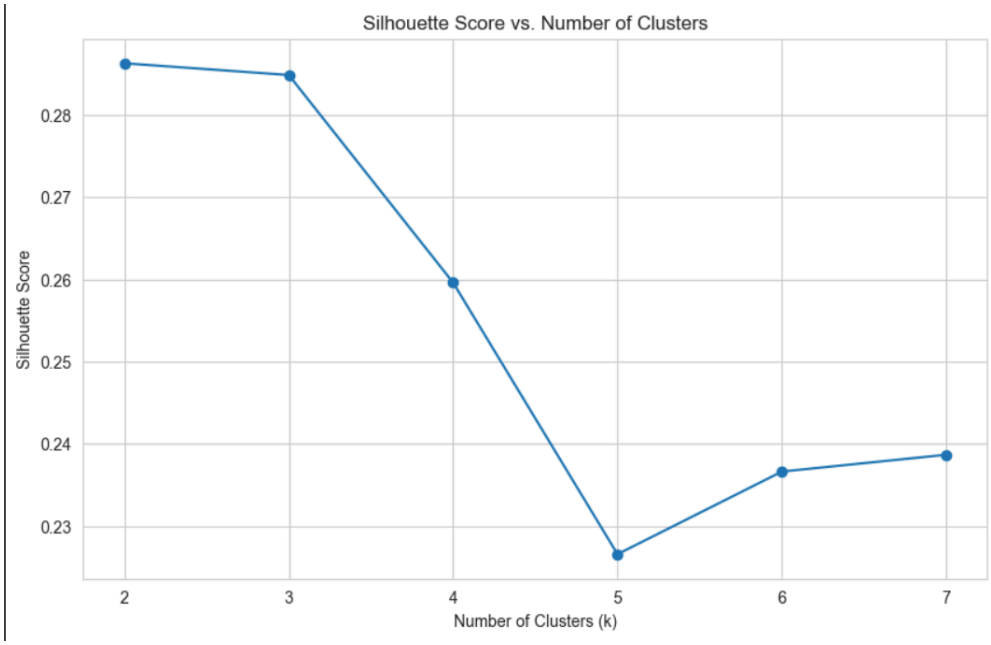
The correlation heatmap revealed a significant correlation between the 'u', 'g', and 'r' features, particularly among those related to photometric measurements across different wavelengths (e.g., PetroRad, petroFlux, petroR50, psfMag). This suggests potential redundancy within the dataset, as multiple features may convey similar information. Conversely, some features exhibited minimal correlation, indicating they might be less useful in predictive models.

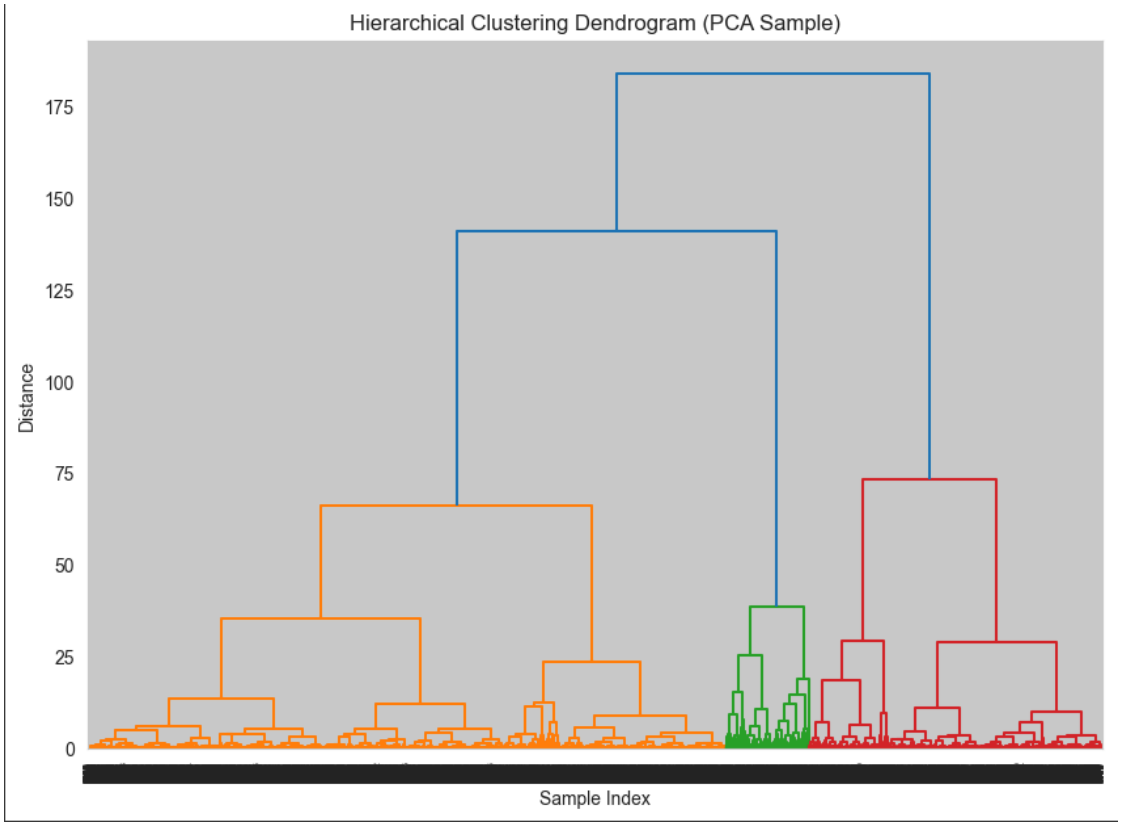


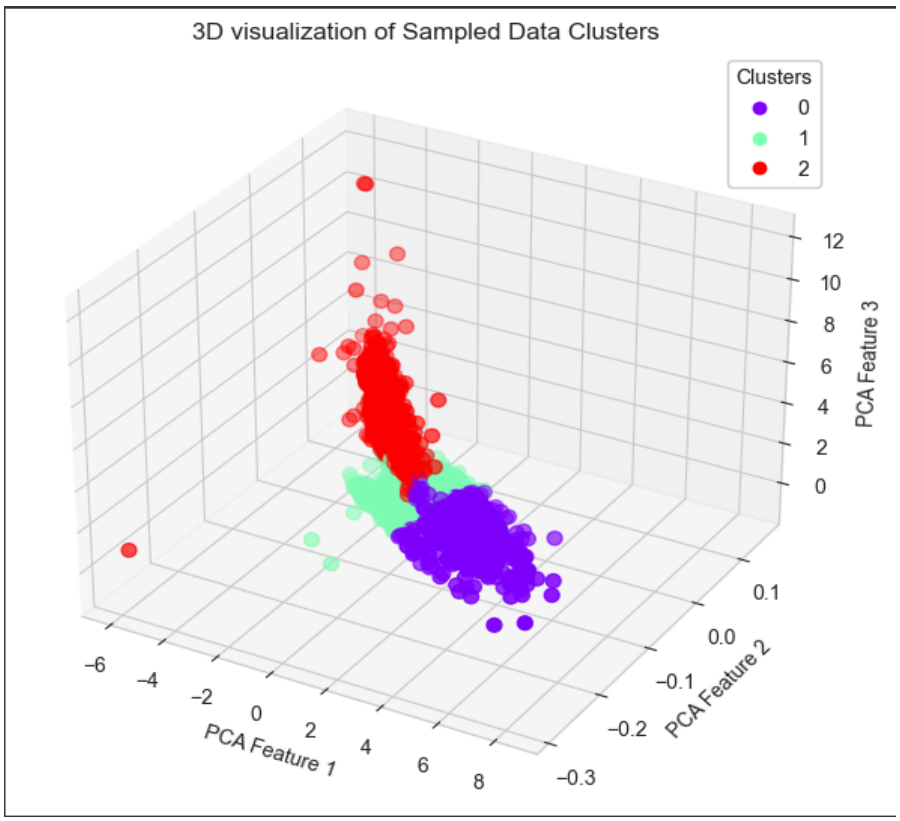


Clustering Analysis

The silhouette score, a measure of cluster separation, was 0.285. This moderate value indicates some level of structure within the clusters, albeit with less distinct separation. This finding prompted an exploration of alternative clustering algorithms, including DBSCAN and hierarchical clustering.




Hierarchical Clustering Dendrogram (PCA Sample)
Distance
175
150
125
100
75
50
25
0
Sample Index


3D visualization of Sampled Data Clusters
Clusters
0
1
2
PCA Feature 1
PCA Feature 2
PCA Feature 3

In hierarchical clustering, we observed three primary clusters (colored orange, green, and red), corresponding to galaxies, stars, and QSOs. The green cluster, representing QSOs, showed a higher degree of similarity within its members. This alignment with the clusters identified through K-means clustering suggests an optimal three-cluster solution for our dataset.

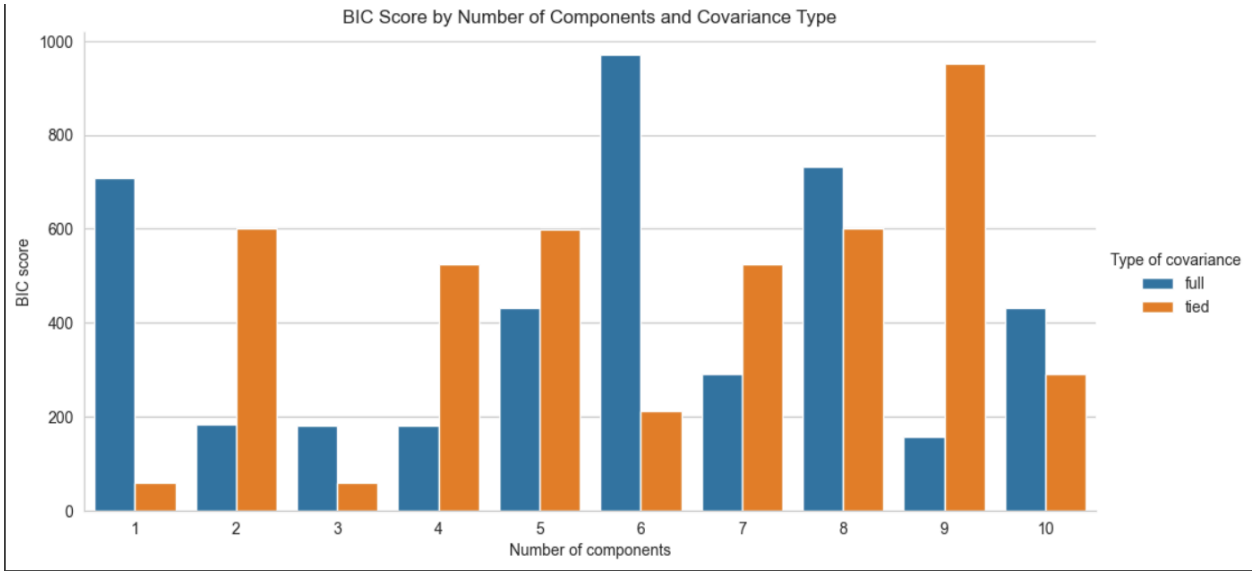
3D Scatter Plots and PCA Analysis

3D scatter plots applying Agglomerative Clustering to PCA-reduced features revealed three distinct clusters, with some overlap between clusters, especially at the boundaries. The dense concentration in the blue cluster indicated high within-cluster similarity, correlating with the green cluster from hierarchical analysis. In contrast, the red cluster appeared more dispersed.

The plot also highlighted the significance of PCA Feature 1 over Features 2 and 3 in distinguishing between clusters, implying its greater role in the clustering process.

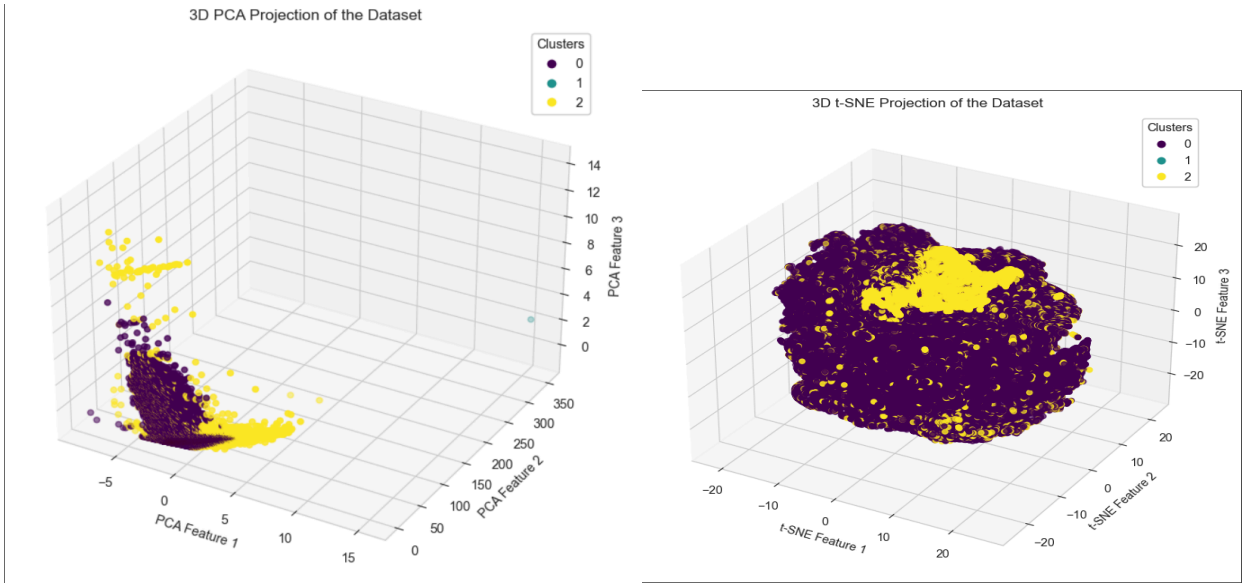
Gaussian Mixture Model (GMM) and Bayesian Information Criterion (BIC)

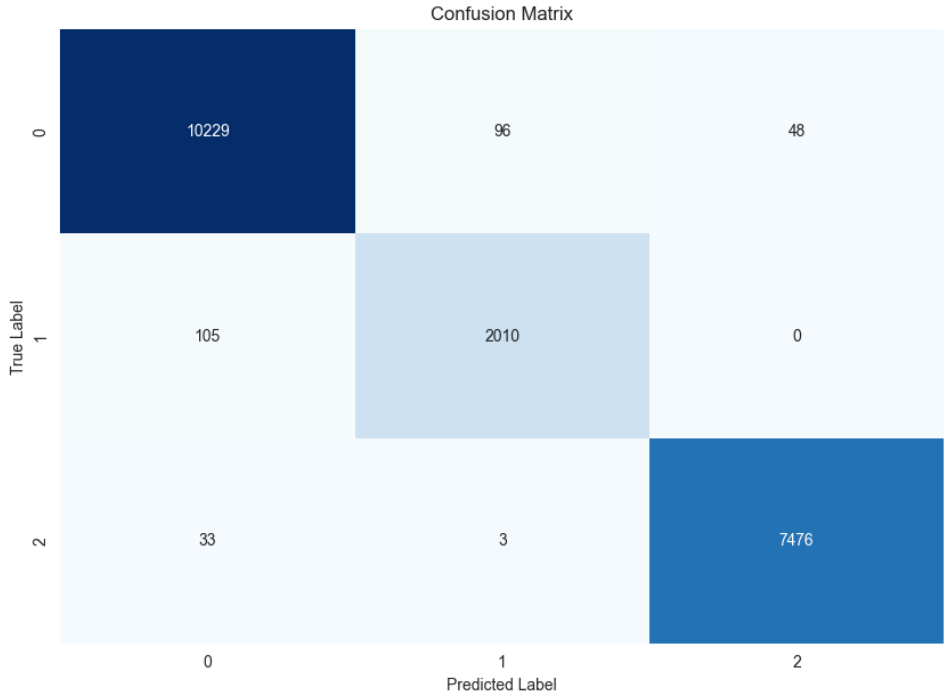
Our analysis of BIC scores for different Gaussian Mixture Models underscored the suitability of a three-component model with tied covariance for our dataset. This model provided the optimal balance between fitting accuracy and complexity, aligning with the assumed three classes of astronomical objects. The lower BIC score for tied covariance models suggested a shared covariance matrix across components was more effective than a full covariance approach.



Principal Component Analysis (PCA)

When reduced to three principal components, the dataset's distribution primarily spanned PCA Features 1 and 2, indicating their greater role in capturing variance and influencing the clustering process. The yellow cluster, in particular, exhibited a more widespread distribution, implying a higher within-cluster variance compared to the others.





Discussion

In the exploration and analysis of astronomical data using machine learning techniques, our project has revealed several key insights that contribute to the broader understanding of

celestial classification and the potential for automated analysis within the field. The discussion below emphasizes our main findings and their significance in the context of astrophysical research.

Our Main Point:

Our model achieved a high level of accuracy (98.58%) in classifying galaxies, stars, and quasars using a Decision Tree Classifier. This result is particularly noteworthy given the inherent challenges in distinguishing between these celestial bodies based on photometric data alone. The success of our approach underscores the power of machine learning techniques in handling complex, high-dimensional datasets and provides a strong foundation for further research and application in automated celestial classification.

Importance of Dimensionality Reduction: The application of PCA and t-SNE significantly enhanced our ability to visualize and interpret the underlying structure of the data. By reducing the dimensionality of the dataset while preserving its intrinsic variance, we were able to identify distinct clusters corresponding to different types of astronomical objects. This not only facilitated a more nuanced analysis but also improved the performance of our clustering algorithms, as evidenced by the meaningful separation achieved in the 3D scatter plots.

Clustering Algorithms Reveal Underlying Structure: The use of various clustering algorithms, including KMeans, Hierarchical Clustering, and Gaussian Mixture Models, proved effective in discerning the natural groupings within the data. The silhouette score and BIC evaluations provided quantitative support for our clustering choices, indicating a good balance between cluster cohesion and separation. This diversity in algorithmic approach allowed us to cross-validate our findings and ensure robustness in our classification strategy.

Potential for Automated Classification in Astronomy: Our work highlights the potential of machine learning methods to revolutionize the field of astronomy by automating the classification of celestial objects. With the volume of data from sky surveys continually increasing, such automated systems can greatly assist in data analysis, freeing up valuable time for astronomers to focus on interpretive and theoretical aspects of their research. Furthermore, the adaptability of our model to different datasets and its scalability makes it a valuable tool for future astronomical studies, especially with the advent of next-generation telescopes.

Challenges and Future Directions: While our results are promising, the moderate silhouette score indicates room for improvement in cluster separation. Future work could explore more advanced or novel machine learning techniques to enhance the distinctiveness of clustering. Additionally, integrating spectral analysis data could further refine the classification accuracy, offering deeper insights into the physical properties of the objects being studied.

In conclusion, our project not only demonstrates the effectiveness of machine learning in classifying astronomical objects but also opens avenues for further exploration and refinement of these techniques. The potential for these methods to contribute to our understanding of the

universe is vast, and we anticipate that continued research will yield even more sophisticated and powerful tools for astronomical analysis.

Ethics & Privacy

In our use of the Sloan Digital Sky Survey (SDSS) dataset for machine learning in astronomy, we encounter ethical and privacy considerations. While astronomical data typically lacks personal identifiers, the risk of inadvertently exposing sensitive information increases when datasets are combined. Additionally, biases inherent in the data could skew model predictions, potentially perpetuating inaccuracies in astronomical research. To address these concerns, we prioritize data privacy and bias mitigation, employing rigorous evaluations and transparent communication of findings. Leveraging tools like Deon (https://deon.drivendata.org), we aim to ensure responsible and ethical use of the SDSS dataset in our astronomical research endeavors through machine learning.

Team Expectations

- *We expect to communicate and evaluate our progress several times a week.*
- *We will relay information through our discord group and github*
- *We expect to handle and disputes civilly, and to distribute work evenly and fairly*
- *We will make a schedule of when we want certain aspects of the project done and adhere to it*
- *We will make sure everyone's opinions are heard*

Conclusion

In this project, we have successfully applied machine learning techniques, including K-Means Clustering, Gaussian Mixture Models (GMMs), and Hierarchical Clustering, along with dimensionality reduction methods like PCA and t-SNE, to effectively classify astronomical objects such as galaxies, stars, and quasars. We achieved a notable accuracy of 98.58% in our Decision Tree Classifier on the test set, showing the effectiveness of our machine learning algorithms. Our findings demonstrate the potential of machine learning in astrophysics, particularly in the classification of celestial bodies.

Our results underline the importance of automated classification systems in handling the vast and complex datasets generated by astronomical technology. The application of our model offers a more efficient and precise categorization of astronomical objects.

Footnotes

1.^: Andrés Almeida et al. "The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V", 2023 ApJS 267 44. https://iopscience.iop.org/article/10.3847/1538-4365/acda98.