# EXXA GSoC 2025 Test Submission – Exoplanet Atmosphere Classification Report

**Table of Contents**



## 1. Introduction

This document summarizes the approach and results for the GSoC 2025 EXXA Test Submission. The goal of the test is to demonstrate the ability to handle FITS images data related to exoplanet atmospheres and also to apply unsupervised learning techniques to cluster the data meaningfully as well derive results.

## 2. Dataset and Problem Statement

The dataset comprises FITS format images representing exoplanet atmosphere data. The challenge involves reading and interpreting this data, extracting meaningful features, and using ML based clustering algorithms to classify these images.

## 3. Environment and Tools Used

- Python
- Google Colab
- Libraries: pandas, numpy, matplotlib, seaborn, sklearn, astropy, joblib.

## 4. Image Preprocessing

- Uploaded FITS image files into Colab.
- Used astropy.io.fits to read the images from the device.
- Normalized and flattened the images to 1D feature arrays for data- processing.

## 5. Feature Extraction

- Extracted numerical features from image data by flattening 2D arrays.
- Applied standardization (StandardScaler) methods to improve clustering performance of ML algorithm.

## 6. Clustering Algorithms Applied

### 6.1 KMeans Clustering

- Used KMeans(n_clusters=3) to group similar image data.
- Worked well with PCA for visualization.

### 6.2 DBSCAN Clustering

- Used density-based clustering to identify more complex clusters.
- Useful for detecting noise and varied density clusters.

## 7. Dimensionality Reduction (PCA)

- Used PCA to reduce high-dimensional image data into 2D space.
- Enabled easy visualization of clusters.

## 8. Visualizations

- Created 2D scatter plots to visualize clusters formed by KMeans and DBSCAN.
- Color-coded clusters for easier interpretation and analysis.

## 9. Model Saving and Loading

- Saved the clustering models using joblib.dump()
- Reloaded the models for reuse using joblib.load()

## 10. Challenges Faced

- Handling and normalizing FITS images in batches.
- Dimensionality reduction for very high-dimensional image data.
- Determining optimal clustering parameters.

## 11. What I Learned

Through this test, I learned:

- How to pre-process scientific FITS images for based ML tasks.
- To extract meaningful features from raw image data using Python.
- The power of dimensionality reduction techniques like PCA.
- How different clustering techniques work and how to evaluate them.
- Efficient model saving and reloading for scalable workflows.

## 12. Conclusion

The test provided valuable hands-on experience in image preprocessing, feature extraction, clustering, and visualization. This task strengthened my skills in ML, scientific computing, and problem-solving, aligning well with the goals of EXXA's GSoC project. If passed I would be happy to futher dive deeper into this project.

## 13. References


- https://docs.astropy.org
- https://scikit-learn.org
- https://joblib.readthedocs.io