BUS6440: Applied Machine Learning for Analytics

Module 4 Assignment

K-Means Python Application

6/28/2025

Taiwo Timothy Babalola

Leaner ID: 162894

Submitted to Dr. Rapheal Wanjiku

# **Overview**

This project explores the application of the K-Means clustering algorithm to mock lunar observation data, simulating the structure of data obtained from the Kaguya Monoscopic Uncontrolled Observations available in the Registry of Open Data on AWS.

## **Purpose**

The goal was to demonstrate how unsupervised machine learning, particularly K-Means clustering, can be used to classify and identify patterns in lunar surface data files, even when detailed labels or domain knowledge are absent.

## **Dataset Description**

Due to access limitations and time constraints, this assignment used simulated .img files, mimicking real lunar imagery metadata. Each file was processed to extract 5 randomly generated features representing surface characteristics. In total, 5 files were generated to perform clustering.

# **Process Summary (via https://github.com/AmazingTaiwo/BAN6440.git)**

* Data import and Simulation: Five .img files were created in a local directory to simulate remote sensing data.
* Feature Extraction: Random features were generated for each file, creating a tabular dataset.
* Preprocessing: Data was standardized using Standard Scaler to normalize the feature space.
* Clustering: K-Means was applied with k=3, and the data was grouped into clusters based on feature similarity.
* Visualization: PCA was used to reduce feature dimensions and visualize the clusters in a 2D scatter plot with file name: kmeans\_pca\_plot

# **Testing and Validation (via: https://github.com/AmazingTaiwo/BAN6440.git)**

Unit tests were implemented to validate data extraction, clustering robustness, and edge cases such as handling too few samples or missing directories. All tests passed, confirming the application’s reliability and generate text file with name: unit\_test\_results

# **Results & Insights**

The K-Means model successfully grouped the mock lunar files into 3 distinct clusters. While the dataset was artificial, the results demonstrate the potential for applying clustering techniques to real-world spatial datasets for pattern detection and anomaly analysis.

# **Conclusion**

This exercise showcases the feasibility of using machine learning with open space datasets. With access to real image features, the same pipeline can support advanced lunar mapping, aid in mission planning, or assist in identifying previously unknown surface structures.