**Module 4: K-Means Python Application**

Richard O. Adeitan

Master of Business Administration, Nexford University

BAN6440: Applied Machine Learning for Analytics

Professor Bucciarelli

1 July 2024

**Executive Summary**

In this project, I developed a K-Means clustering application using a dataset named "segmentation data.csv". The primary goal was to perform meaningful clustering on the dataset, utilizing Python and various machine learning libraries. The entire process involved data preprocessing, implementing the K-Means algorithm, conducting unit tests, and performing an analysis of the results.

**Dataset and Preprocessing**

The dataset "segmentation data.csv" was used for this project. It contains various features that required standardization before applying the K-Means algorithm. StandardScaler from the sklearn library was employed to scale the data, ensuring that all features contributed equally to the clustering process.

**K-Means Application**

I implemented the K-Means algorithm in a Python script named kmean\_application.py. The script:

* Loaded and preprocessed the dataset.
* Applied the K-Means clustering algorithm to identify three distinct clusters.
* Saved the clustered data into a new file named segmentation\_data\_with\_clusters.csv.
* Visualized the clusters to provide an intuitive understanding of the segmentation.

**Unit Testing**

To ensure the robustness and reliability of the K-Means implementation, I developed a separate unit testing script named test\_kmeans.py. This script:

* Loaded and standardized the dataset.
* Applied the K-Means clustering algorithm.
* Verified the number of clusters and the silhouette and Davies-Bouldin scores to assess clustering quality.

**Analysis and Insights**

The visualization of the clusters revealed distinct groupings within the data, indicating successful segmentation. Evaluation metrics such as silhouette score and Davies-Bouldin index confirmed the quality of the clustering. These metrics provided insights into the compactness and separation of the clusters, validating the effectiveness of the K-Means algorithm in segmenting the dataset.

**Challenges and Solutions**

One of the primary challenges was ensuring data integrity during preprocessing. Exception handling was incorporated to manage potential errors, such as file reading issues and data scaling problems. Additionally, thorough unit testing helped identify and resolve any bugs in the clustering implementation.

**Conclusion**

This project successfully demonstrated the application of the K-Means algorithm to segment a dataset. The detailed implementation and rigorous testing ensured the reliability of the results. The clustered data can be further analyzed to derive actionable insights, making it a valuable tool for data-driven decision-making.