Course Final Project : Unsupervised Machine Learning (Clustering)

Student Dropout and Academic Success

**Objective of the Analysis**

The main objective of this project is to develop a machine learning model capable of clustering into one of three categories. Original dataset has three different target class: Dropout, Enrolled, or Graduate. The goal is to explore various clustering techniques and analyze their effectiveness in identifying patterns within the student data. While the original dataset includes a predefined target variable, it is only used for evaluation rather than training the clustering models.

**Dataset Overview**

The dataset was acquired from the UCI Machine Learning Repository, created from a higher education institution's student records. The data was gathered from several disjoint databases related to students enrolled in different undergraduate degrees. This dataset aims to help predict student dropout and academic success. It help identify at-risk students, ultimately contributing to efforts aimed at reducing academic dropout and failure in higher education. This dataset consist :

* Number of Instances: 4424 student records
* Number of Features: 36, covering a wide range of attributes such as academic path, demographics, and socio-economic factors.
* Target/Predictor: Three category classification (dropout, enrolled, and graduate) representing student risk at the end of the normal duration of the course. This column will not be used in clustering model. However will used as reference to compared with model result

**Data Exploration and Preprocessing**

**Exploratory Data Analysis (EDA)**

* No missing values were found in the dataset.
* Feature correlations were examined, revealing some highly correlated columns.
* All features were numeric, which streamlined the clustering process.

**Feature Engineering and Transformation**

* Most columns were integers, so they were converted to float data types for consistency.
* Some features exhibited high skewness, which was corrected using logarithmic transformation.
* Standard Scaling was applied to normalize the data, ensuring all features were on a comparable scale.
* Stratified Shuffle Split was implemented to maintain the class distribution when splitting data for evaluation.

**Model Training**

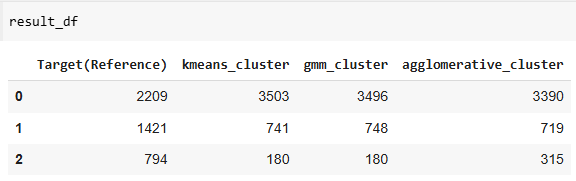
Three different unsupervised machine learning models used. :

1. KMeans
2. Gaussian Mixture
3. Agglomerative Clustering

The same dataset was used for all three models, and the predicted cluster labels were appended to the dataset. To compare model performance, value counts for each model’s predicted clusters and the original target labels were stored in a separate DataFrame for evaluation.

**Model Evaluation**

All three clustering models produced similar results, indicating that the dataset contains clear structure and patterns. K-Means and GMM showed particularly close results, suggesting that both approaches capture similar underlying distributions in the data. When comparing model predictions to the target column, discrepancies were observed, but a notable pattern emerged where one cluster consistently had a much higher representation compared to the other two.



**Final Model Recommendation**

While the initial clustering models provided valuable insights, there are several ways to enhance model performance:

1. Dimensionality Reduction: Applying Principal Component Analysis (PCA)
2. Hyperparameter Tuning: Adjusting the number of clusters, covariance type (for GMM), and linkage criteria (for Agglomerative Clustering) may improve results.
3. Feature Selection: Removing redundant or highly correlated features could improve model interpretability and efficiency.

This project demonstrates the potential of clustering models to uncover patterns in student success and dropout trends. Further refinements and additional data sources could enhance predictive capabilities, supporting educational institutions in proactive decision-making.