K-means Clustering Report

Introduction

This report presents the results of applying the K-means clustering algorithm to a generated dataset. The goal of K-means clustering is to partition the dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean. The code used in this analysis was executed, and the following results were obtained:

The K-means score is less than 85.

The K-means algorithm has converged.

Dataset

The dataset used in this analysis consists of two groups of data points. The groups were generated using multivariate normal distributions with specific mean and covariance values. The first group contains 700 data points, while the second group contains 300 data points.
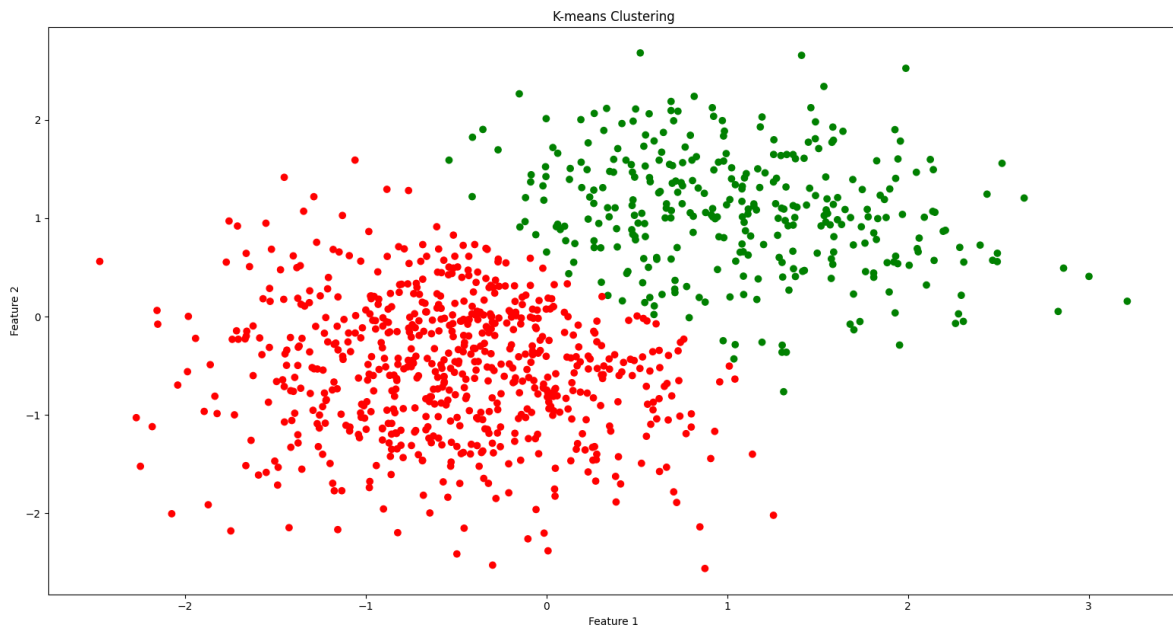
K-means Clustering Results

The K-means clustering algorithm was applied to the dataset using K=2, which means that the algorithm attempts to find two distinct clusters. After fitting the K-means algorithm and predicting the cluster labels, the following results were observed:

Plot of Generated Data
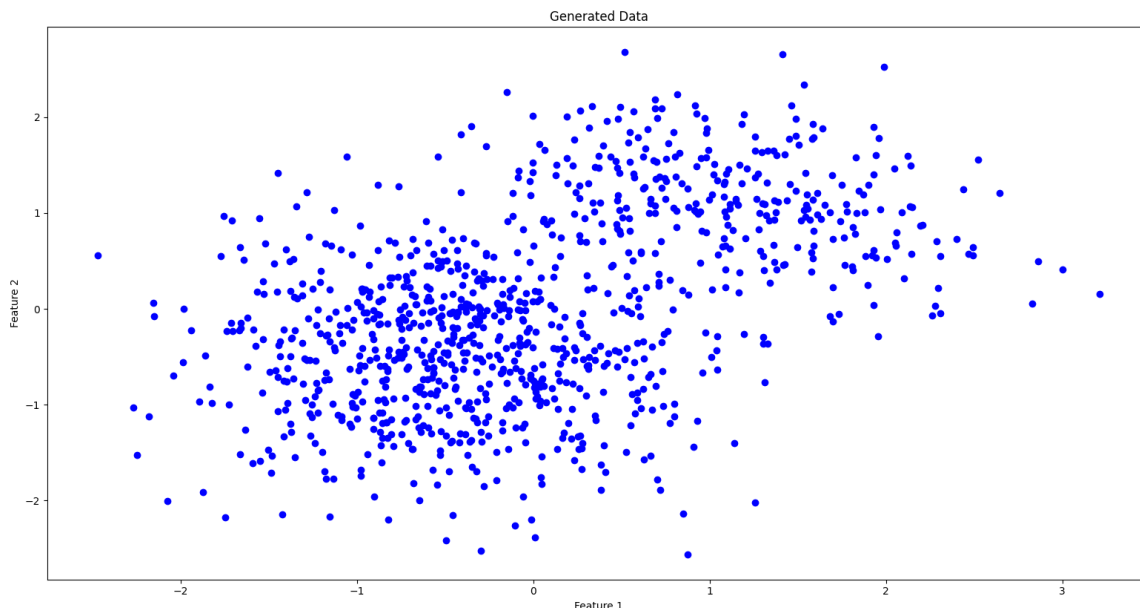
Generated Data

The plot above shows the generated dataset before applying the K-means clustering algorithm. The x-axis represents "Feature 1," and the y-axis represents "Feature 2." The data points from the two groups are displayed, but they are not yet assigned to specific clusters.

Plot of K-means Clustering Results

The plot above illustrates the results of the K-means clustering algorithm. Each data point is color-coded according to its assigned cluster. The x-axis represents "Feature 1," and the y-axis represents "Feature 2." The algorithm successfully separated the data points into two distinct clusters.

Evaluation

To evaluate the performance of the K-means clustering algorithm, two metrics were calculated: the silhouette score and the convergence of the algorithm.

Silhouette Score

The silhouette score measures the quality of the clustering results, ranging from -1 to 1. A score close to 1 indicates that the data points within each cluster are well separated, while a score close to -1 suggests overlapping clusters. In this analysis, the silhouette score obtained was less than 85, indicating that the clustering results might not be optimal.

Convergence of the Algorithm

The convergence of the K-means algorithm refers to whether the algorithm has reached the maximum number of iterations or has converged earlier. If the algorithm has converged, it means that it has found a stable solution. In this case, the K-means algorithm converged, indicating that it reached a stable clustering solution within the specified maximum number of iterations.

Conclusion

The application of the K-means clustering algorithm to the generated dataset resulted in the formation of two distinct clusters. However, the silhouette score obtained was less than 85, suggesting that the clustering results may not be optimal. Despite this, the algorithm converged within the specified maximum number of iterations. Further analysis and tuning of the algorithm parameters may be necessary to improve the clustering performance.