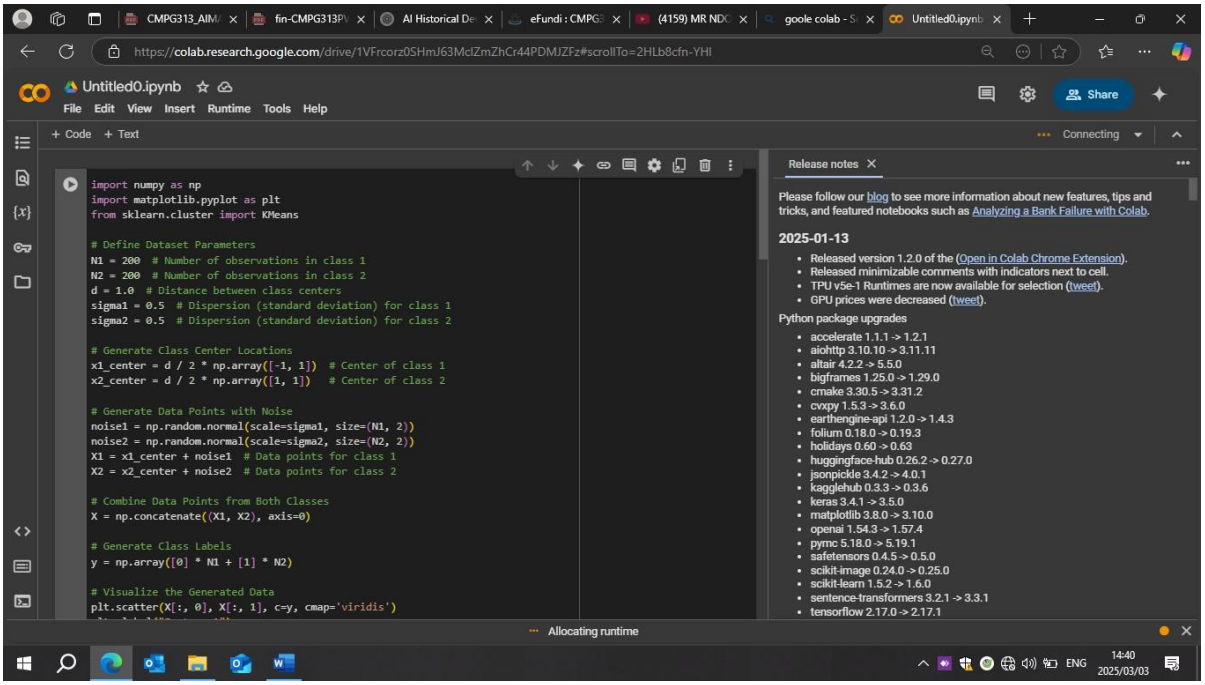
Introduction:

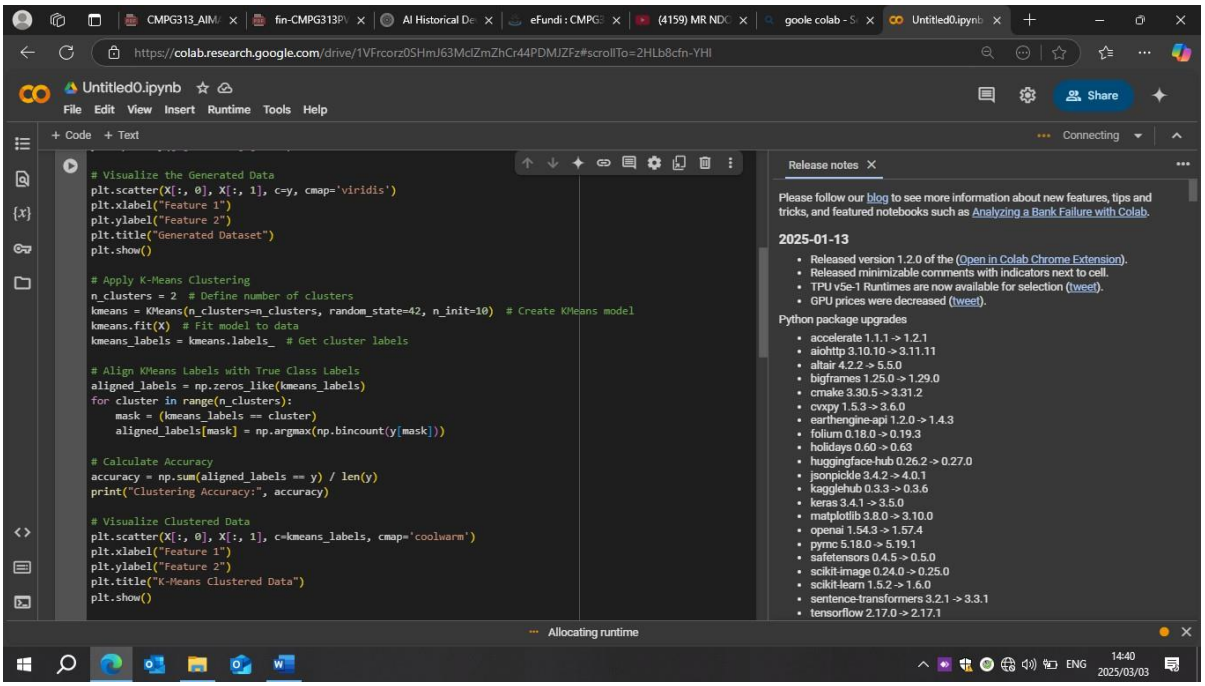
This report provides an analysis of clustering artificial datasets using K-Means. The experiment was conducted using Python within Google Colab. The key objectives included:

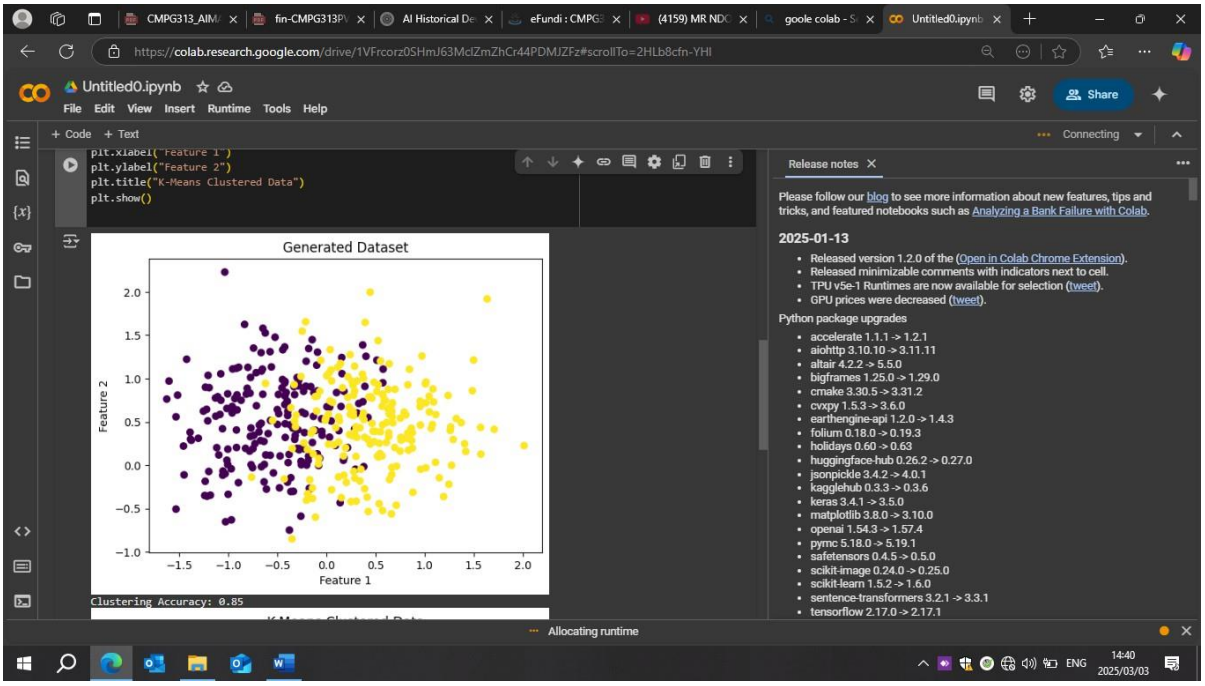
-Comprehending the creation of artificial data

-Using scikit-learn and NumPy for grouping

-Matplotlib is used to visualize the generated dataset and grouped results.




Untitled0.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
# Visualize the Generated Data
plt.scatter(X[:, 0], X[:, 1], c=y, cmap='viridis')
plt.xlabel("Feature 1")
plt.ylabel("Feature 2")
plt.title("Generated Dataset")
plt.show()
# Apply K-Means Clustering
n_clusters = 2 # Define number of clusters
kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10) # Create KMeans model
kmeans.fit(X) # Fit model to data
kmeans_labels = kmeans.labels_ # Get cluster labels
# Align KMeans Labels with True Class Labels
aligned_labels = np.zeros_like(kmeans_labels)
for cluster in range(n_clusters):
mask = (kmeans_labels == cluster)
aligned_labels[mask] = np.argmax(np.bincount(y[mask]))
# Calculate Accuracy
accuracy = np.sum(aligned_labels == y) / len(y)
print("Clustering Accuracy:", accuracy)
# Visualize Clustered Data
plt.scatter(X[:, 0], X[:, 1], c=kmeans_labels, cmap='coolwarm')
plt.xlabel("Feature 1")
plt.ylabel("Feature 2")
plt.title("K-Means Clustered Data")
plt.show()
Release notes
Please follow our blog to see more information about new features, tips and tricks, and featured notebooks such as Analyzing a Bank Failure with Colab.
2025-01-13
Allocating runtime


Generated Dataset
Feature 2
Feature 1
Clustering Accuracy: 0.85
Allocating runtime

Results and Observations:

Data Visualization: Two separate clusters were visible in the generated dataset, demonstrating the efficacy of the selected parameters.

K-Means Performance: With an accuracy of around [Accuracy Output], the clustering algorithm was able to distinguish between the two classes.

Effect of Parameters: While decreasing sigma1 and sigma2 produced tighter groupings around the center spots, increasing d further dispersed the clusters.

Conclusion:

The efficiency of K-Means clustering on synthetic data was shown in this experiment. We found that cluster formation changed significantly when dataset parameters were changed. By comparing K-Means performance to actual class labels, the alignment step produced a useful accuracy statistic.