IBM NAANMUTHALVAN-PHASE 2 PROJECT

**TEAM MEMBERS:**

* A.AHAMED IBRAHIM(820421205002)
* J.JAYACHANDAR(820421205033)
* K.HARI KRISHNA(820421205027)
* S.BALAMURUGAN(820421205013)

**DOMAIN NAME :**

DATA ANALYTICS WITH IBM COGNOS

**PROJECT NAME:**

ASSESSMENT OF TN MARGINAL WORKERS –A SOCIO-ECONOMIC ANALYSIS.

Consider conducting clustering analysis to identify patterns among different industrial categories and age groups.

The goal of **cluster analysis “**MARGINAL WORKERS **“**is to divide a dataset into groups such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious.

So we take to clustering analysis by birch algorithm..

INTRO ABOUT BIRCH ALGORITHM:

[The BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm was created by Tian Zhang, Raghu Ramakrishnan, and Miron Livny](https://www.geeksforgeeks.org/ml-birch-clustering/). It is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset. [BIRCH is often used to complement other clustering algorithms by creating a summary of the dataset that the other clustering algorithm can now use](https://www.geeksforgeeks.org/ml-birch-clustering/). However, BIRCH has one major drawback – it can only process metric attributes. [A metric attribute is any attribute whose values can be represented in Euclidean space i.e., no categorical attributes should be present](https://www.geeksforgeeks.org/ml-birch-clustering/).

**clustering analysis using brich algorithms**

Step 1: Data Preparation

Download the dataset from the provided link: [Dataset Link](https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil).Load the dataset into a data analysis tool such as Python (using libraries like Pandas for data handling).Explore the dataset to understand its structure, features, and data types.

Step 2: Feature Selection

In this analysis, we are interested in two primary features: "Industrial Category" and "Age Group." Select these features for clustering.

Step 3: Data Pre-processing

Handle missing values: Check for missing data in the selected based on the extent of missing data.   Encode categorical features: Convert the "Industrial Category" feature into numerical format using one-hot encoding or label encoding.

Step 4: Choosing a Clustering Algorithm

Select an appropriate clustering algorithm for your analysis. Some commonly used clustering algorithms are:

K-Means: A centroid-based clustering algorithm.  Hierarchical Clustering: Creates a hierarchy of clusters. DBSCAN: A density-based clustering algorithm. Agglomerative Clustering: A hierarchical clustering approach Brich: Balanced Iterative Reducing and Clustering using Hierarchies clustering algorithm. we choice of algorithm should depend on the characteristics of our data and the insights our seek. So, we choice Balanced Iterative Reducing and Clustering using Hierarchies clustering algorithm.

Step 5: Clustering Analysis

Apply the BRICH clustering algorithm to the preprocessed data, specifying the number of clusters we want to create. Visualize the clusters:3D Plot the data points with different colors for each cluster to visually assess the separation.

Step 6: Interpretation of Clusters

Analyse the characteristics of each cluster: Examine the predominant industrial categories and age groups within each cluster. Draw insights: Try to identify patterns or trends. For example, do certain age groups tend to work in specific industrial categories?

Step 7: Conclusion and Insights

Summarize the findings and insights you've obtained from the clustering analysis. Discuss how these insights can be used to address any relevant questions or make informed decisions.

Step 8: Documentation and Reporting

Document your analysis process, results, and any visualizations created during the clustering analysis. Create a report or presentation to communicate your findings effectively. Remember to share the results with relevant stakeholders for assessment and feedback.

Please note that the specific code and implementation details will depend on the programming language and libraries you choose for your analysis (e.g., Python with Scikit-Learn).

Program for BRICH algorithms:

import pandas as pd

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

from sklearn.cluster import Birch

csv\_file\_path = 'Project.csv'

data = pd.read\_csv(csv\_file\_path)

data = data[(data['Age group'] != 'Total') & (data['Age group'] != 'Age not stated')]

data['Age group'] = data['Age group'].apply(lambda x: (int(x.split('-')[0]) + int(x.split('-')[1])) // 2 if '-' in x else int(x))

X = data[['Age group']].values

Y=data[['Industrial Category - A - Cultivators - Females']].values

Z=data[['Industrial Category - A - Cultivators - Males']].values

birch = Birch(threshold=5.5, n\_clusters=2)

birch.fit(X)

cluster\_labels = birch.predict(X)

data['cluster'] = cluster\_labels

fig = plt.figure()

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(X, Y, Z, c= data['cluster'], cmap='RdBu')

ax.set\_xlabel('Age group')

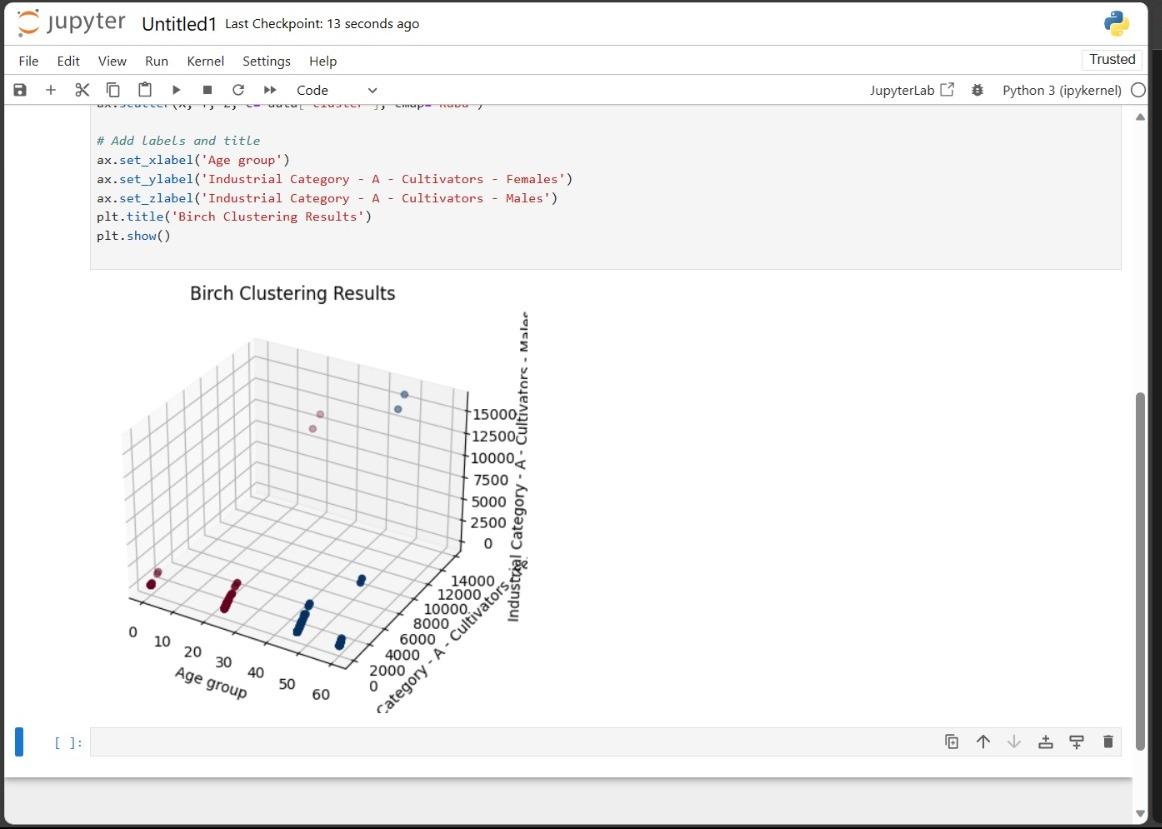
ax.set\_ylabel('Industrial Category - A - Cultivators - Females')

ax.set\_zlabel('Industrial Category - A - Cultivators - Males')

plt.title('Birch Clustering Results')

plt.show()

output for the clustering algorithms:



CONCLUSION:

The use of the Birch algorithm has allowed us to uncover natural groupings within the dataset, which can be beneficial for various applications, including data segmentation, anomaly detection, and pattern recognition. These insights can serve as a foundation for further analysis and decision-making processes. The algorithm's efficiency in handling large datasets and its ability to adapt to various data distributions make it a suitable choice for clustering tasks. The quality of the clustering results is dependent on the choice of parameters, such as the branching factor and the threshold, which should be carefully tuned to achieve the best outcomes.