**ASSESSMENT OF MARGINAL WORKERS IN TAMIL NADU.**

**A SOCIOECONOMICAN ANALYSIS (ADS).**

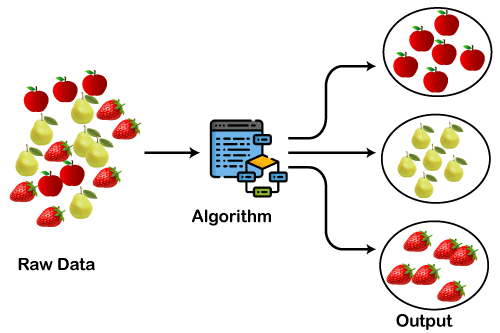
## PROJECT TOPIC : Assessment of Marginal Workers in TamilNadu.



This project is conducting clustering analysis to identify patterns among different industrial categories and age groups.

**Cluster analysis**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more [similar](https://en.wikipedia.org/wiki/Similarity_measure) (in some sense) to each other than to those in other groups (clusters). It is a main task of [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics), [data compression](https://en.wikipedia.org/wiki/Data_compression), [computer graphics](https://en.wikipedia.org/wiki/Computer_graphics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning).



**TYPES OF CLUSTERING** :

* The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist.
* Below are the main clustering methods used in Machine learning:

1. **Partitioning Clustering**
2. **Density-Based Clustering**
3. **Distribution Model-Based Clustering**
4. **Hierarchical Clustering**
5. **Fuzzy Clustering**

**Clustering Algorithms**

The Clustering algorithms can be divided based on their models that are explained above. There are different types of clustering algorithms published, but only a few are commonly used. The clustering algorithm is based on the kind of data that we are using. Such as, some algorithms need to guess the number of clusters in the given dataset, whereas some are required to find the minimum distance between the observation of the dataset.

Here we are discussing mainly popular Clustering algorithms that are widely used in machine learning:

1. **K-Means algorithm:** The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of **O(n).**
2. **Mean-shift algorithm:** Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model, that works on updating the candidates for centroid to be the center of the points within a given region.
3. **DBSCAN Algorithm:** It stands **for Density-Based Spatial Clustering of Applications with Noise**. It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages. In this algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.
4. **Expectation-Maximization Clustering using GMM:** This algorithm can be used as an alternative for the k-means algorithm or for those cases where K-means can be failed. In GMM, it is assumed that the data points are Gaussian distributed.
5. **Agglomerative Hierarchical algorithm:** The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure.
6. **Affinity Propagation:** It is different from other clustering algorithms as it does not require to specify the number of clusters. In this, each data point sends a message between the pair of data points until convergence. It has O(N2T) time complexity, which is the main drawback of this algorithm.

**Applications of Clustering**

Below are some commonly known applications of clustering technique in Machine Learning:

* **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
* **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.
* **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
* **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
* **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

**Program:**

import pandas as pd

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

# Load your dataset

data = pd.read\_csv('your\_dataset.csv')

# Select relevant features

X = data[['industrial\_category', 'age']]

# Choose the number of clusters (k)

k = 3

# Apply K-Means clustering

kmeans = KMeans(n\_clusters=k)

data['cluster'] = kmeans.fit\_predict(X)

# Visualize the results

plt.scatter(data['industrial\_category'], data['age'], c=data['cluster'])

plt.xlabel('Industrial Category')

plt.ylabel('Age')

plt.title('Clustering Results')

plt.show()

|  |  |  |
| --- | --- | --- |
| Industrial Category | Age | Clustering Results |
| LABOURS | 51 | 32 |
| WORKERS | 55 | 99 |

OUTPUT: