Graphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Student Name

***Sanatkumar Rajmogali Ippalpalli***

Title of Project Report

***Guided Project 7 – Spherical KMeans Clustering***

Table of Contents

[Introduction 4](#_Toc83570261)

[Eckovation theme & Question 5](#_Toc83570262)

[Prerequisites before starting coding 6](#_Toc83570263)

[program DEVELOPMENT steps 8](#_Toc83570264)

[Dataset/Image requirements 8](#_Toc83570265)

[Technique – KMeans clustering algorithm 8](#_Toc83570266)

[PROGRAM / CODE DEVELOPMENT 9](#_Toc83570267)

[Analysis 14](#_Toc83570268)

[CONCLUSION 16](#_Toc83570269)

[Figure 1 Import libraries and datasets/modules 9](#_Toc83570291)

[Figure 2 Load Datasets 9](#_Toc83570292)

[Figure 3 Visualization of Image 10](#_Toc83570293)

[Figure 4 Code for estimate number of optimum clusters 10](#_Toc83570294)

[Figure 5 Model Building to Form clusters plots 11](#_Toc83570295)

[Figure 6 Confusion Matrix 14](#_Toc83570296)

***EXECUTIVE SUMMARY***

Science and technology improved many technologies and has guided numerous innovative features which advanced image processing technology.

Clustering is an unsupervised task in machine learning. K-means clustering is a simple but powerful method of clustering method which is based on a centroid-based technique. We need to define the value of k before going with clustering. Among others, the Elbow method is easy to implement to find the best value of k which calculates the WCSS for each value of k to find the suitable value of k. The selection of the value of k is a crucial step in clustering with k-means clustering.

There are different methods, and one of the most popular methods is the **k-means clustering algorithm.** K-Means clustering algorithm is an unsupervised algorithm, and it is used to segment the interest area from the background. The classical k-means method of clustering minimizes the sum of squared distances between cluster centres and cluster members. The intuition is that the radial distance from the Cluster- Centre should be similar for all elements of that cluster. The spherical k-means algorithm, however, is equivalent to the k-means algorithm with cosine similarity, a popular method for clustering high-dimensional data. The idea is to set the centre of each cluster such that it makes the angle between components both uniform and minimal. KMeans clustering is a fundamental tool in machine learning algorithms.

As part of guided project, one of the technologies at identifying spherical KMeans **clustering** as the details mentioned in the question.

# Introduction

Science and technology improved many technologies and has guided numerous innovative features which advanced the techniques in deep learning impacting computer vision, image processing.

Clustering is a set of techniques used to partition data into groups, or clusters. Clusters are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. In practice, clustering helps identify two qualities of data:

* Meaningfulness
* Usefulness

We can perform clustering using many different approaches—so many, in fact, that there are entire categories of clustering algorithms. Each of these categories has its own unique strengths and weaknesses.

The *k***-means clustering** method is an [unsupervised machine learning](https://en.wikipedia.org/wiki/Unsupervised_learning) technique used to identify clusters of data objects in a dataset. There are many different types of clustering methods, but *k*-means is one of the oldest and most approachable. These traits make implementing *k*-means clustering in Python reasonably straightforward, even for novice programmers and data scientists.

Hence Eckovation includes this guided project in the courseware for students to understand, implementation / execute the code themselves.

This report includes the 5W1H about the theme of development of code and running the code with database available over the internet. At the end of the report, the conclusions share the adaptive thresholding & OTSU thresholding features extracted and useful for next course of activities to gain advantages in the edge detection activities development.

# Eckovation theme & Question

**Theme : K-Means Clustering: Image Segmentation**

The classical k-means method of clustering minimizes the sum of squared distances between cluster centres and cluster members. The intuition is that the radial distance from the

Cluster- Centre should be similar for all elements of that cluster. The spherical k-means algorithm, however, is equivalent to the k-means algorithm with cosine similarity, a popular method for clustering high-dimensional data. The idea is to set the centre of each cluster such that it makes the angle between components both uniform and minimal.

**Question:**

Generate a dummy dataset using Scikit-Learn having high dimensionality (number of features >10) and total 4 classes. For this dataset, first implement K-Means clustering and then use the clusters for classification purpose. Now using the same dataset, implement spherical clustering and then check accuracy for classification. Notice the change in accuracy. You may also plot the obtained clusters from both the methods using t-SNE plots or by projecting data into two dimensions using PCA.

# Prerequisites before starting coding

1. Who - Software needed?
2. What - Version / Release of software?
3. Any Prerequisites
4. How - to install the software
5. Which -libraries are needed to execute the problem statement
6. Where – dataset requirements, path location to include in the code
7. When – to use the above feature extraction
8. Who – Software neeed?

Python

1. What- Version / Release of software?

Python version 3.6 (latest version of python)

1. Any Prerequisites

RAM space availability & hard disk space availability

Admin rights to install the software

1. How - to install the software
2. The following url <https://www.python.org/downloads/>can be referred to download python.
3. Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/>
4. Which -libraries are needed to execute the problem statement
5. Numpy (pip install numpy)
6. Matplotlib (pip install matplotlib)
7. Pandas
8. Seaborn
9. Sklearn datasets
10. Sklearn metrics
11. Sklearn clusters
12. Where – dataset requirements, path location to include in the code
13. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: [https://www.pythoncentral.io/add-python-to-path-python-is-not- recognized-as-an-internal-or-external-](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/) [command/](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/).
14. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.
15. When – to use the above feature extraction
16. When – to use the above technique

There are many other [applications of clustering](https://en.wikipedia.org/wiki/Cluster_analysis#Applications), such as document clustering and social network analysis. These applications are relevant in nearly every industry, making clustering a valuable skill for professionals working with data in any field.

# program DEVELOPMENT steps

* Dataset/Image requirement
* Technique selections
* Program / code development
* Analysis

### Dataset/Image requirements

The image sources used for this project are:

Make\_blobs

### Technique – KMeans clustering algorithm

Conventional *k*-means requires only a few steps.

The first step is to randomly select *k* centroids, where *k* is equal to the number of clusters you choose. **Centroids** are data points representing the center of a cluster.

Text

Description automatically generated

Let us hop to the inscribing carving!

### PROGRAM / CODE DEVELOPMENT

As explained step by step during the lecture by mentor, we would approach steps and understand the basics with brief explanation as needed.

#### Step 1: Import the relevant libraries and applicable datasets/modules

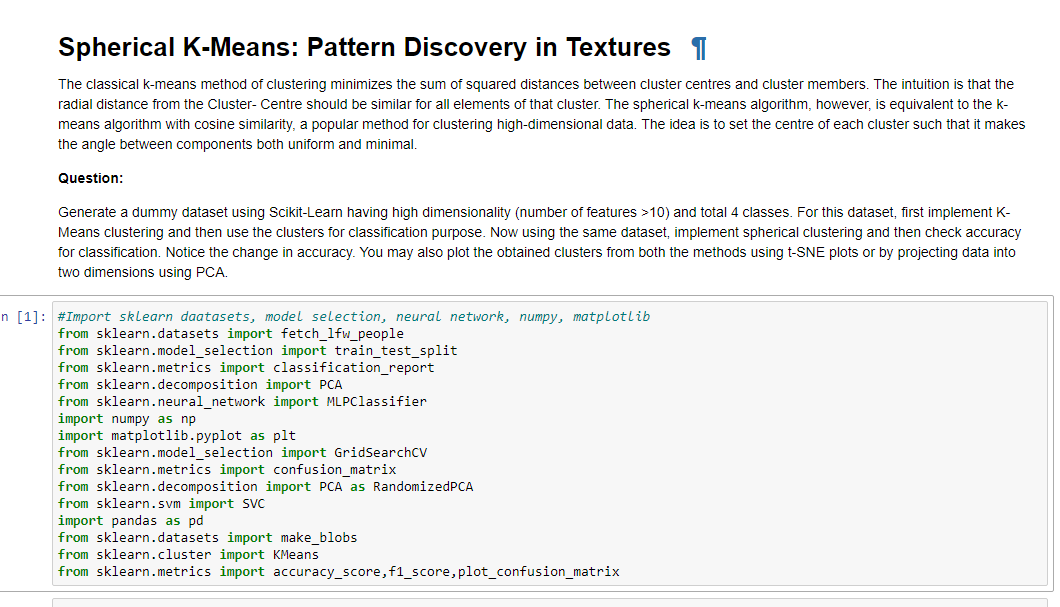


Figure Import libraries and datasets/modules

Step 2: Load dataset/Image and convert Pandas Dataframe

Graphical user interface, text, application

Description automatically generated

Figure Load Datasets

Visualize Images

Chart, scatter chart

Description automatically generated

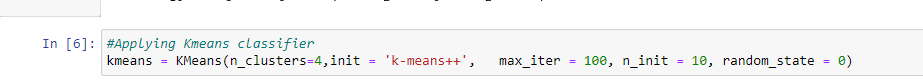
Figure Visualization of Image

#### Step 3: Estimate optimum number of clusters

Figure Code for estimate number of optimum clusters

#### Step 4: Model Building

Apply the KMeans algorithm to build model



A picture containing scatter chart

Description automatically generated

Bubble chart

Description automatically generated

Table

Description automatically generated with medium confidence

Figure Model Building to Form clusters plots

Here are the parameters used in this example:

* **init** controls the initialization technique. The standard version of the *k*-means algorithm is implemented by setting init to "random". Setting this to "k-means++" employs an advanced trick to speed up convergence, which you’ll use later.
* **n\_clusters** sets *k* for the clustering step. This is the most important parameter for *k*-means.
* **n\_init** sets the number of initializations to perform. This is important because two runs can converge on different cluster assignments. The default behavior for the scikit-learn algorithm is to perform ten *k*-means runs and return the results of the one with the lowest SSE.
* **max\_iter** sets the number of maximum iterations for each initialization of the *k*-means algorithm.

#### Step 5: Image Segmentation

There are different methods, and one of the most popular methods is the k-means clustering algorithm. K-Means clustering algorithm is an unsupervised algorithm, and it is used to segment the interest area from the background.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, Word, chat or text message

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Background pattern

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

Table

Description automatically generated

Figure Confusion Matrix

This is continued to achieve desired results as per requirements.

### Analysis

The KMeans unsupervised learning clustering technique is completed.

We can use the techniques you learned here to cluster your own data, understand how to get the best clustering results, and share insights with others.

This entire program runs within few seconds.

# CONCLUSION

In this guided project, we built Spherical KMeans algorithm model.

This is done in first attempt. Hence, the improvements in the code with time with multiple attempts may be checked and justified for the accuracy score.

This entire program runs within few seconds.

references:

1. <https://datatofish.com/k-means-clustering-python/>
2. <https://codefires.com/implementation-of-k-means-clustering-in-python/>
3. <https://realpython.com/k-means-clustering-python/>