	Hope Foundation's **Finolex Academy of Management and Technology, Ratnagiri** **Department of Information Technology**		
Subject name	**Business Intelligence Lab**		**Subject Code: ITL602**
Class	**TE IT**	**Semester – VI (CBCGS)**	**Academic year: 2018-19 (FH 2019)**
Name of Student			**QUIZ Score :**
Roll No		**Assignment/Experiment No:**	**06**
Title:	**Implementation of k-means clustering using Java.**		

1. Course objectives applicable: LO4

2. Course outcomes applicable: LO4

3. Learning Objectives:

1. To learn K-means clustering method.
2. To learn the iterations of k-means algorithm and data distribution.
3. To develop a program to implement k-means algorithm.

4. Practical applications of the assignment/experiment: clustering of data in data analysis.

5. Prerequisites:

1. K-means algorithm theoretically
2. JDK 8.0 / Python 3.6

6. Hardware Requirements:

1. PC with minimum 2 GB RAM

7. Software Requirements:

1. Windows / Linux
2. JDK 8.0 / Python 3.6

8. Viva Questions (if any): (Online Quiz will be taken separately batch-wise)

1. How the distance between objects is calculated?
2. In K-means algorithm what is denoted by 'K'?
3. What is within cluster variation? How it relates with cluster analysis?
4. How to relate inter-cluster and intra-cluster distance with clustering quality?

9. Experiment/Assignment Evaluation:

Sr. No.	Parameters	Marks obtained	Out of
1	Technical Understanding (Assessment may be done based on Q & A **or** any other relevant method.)		6
2	Neatness/presentation		2
3	Punctuality		2

Date of performance (DOP)		**Total marks obtained**		**10**
Date of checking (DOC)		**Signature of teacher**		

10. Theory: **<<handwritten work>>**

***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm.**

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

Initialization methods:

Commonly used initialization methods are Forgy and Random Partition. The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while Random Partition places all of them close to the center of the data set.

Solve the following using k-Means algorithm

Data = {3, 7, 1, 9, 8, 15, 5, 26, 12, 15}

Number of clusters, k=3

11. Performance Steps:

1. Implement a Java/Python program to perform clustering on the given data.
2. Run the program to display the elements cluster-wise.

12. Results:

<< Add the hard-copy of source code and the output screen shots >>

13. Learning Outcomes Achieved

1. Understanding of k-means algorithm
2. Understanding of iterative methods of distribution of cluster elements.
3. Understanding the program implementation of k-means.

14. Conclusion:

1. **Applications of the studied technique in industry:** In real-time data analysis.
2. **Engineering Relevance:** To preform clustering of data for data analysis.
3. **Skills Developed:** Program implantation for k-means algorithm.

15. References:

[1] Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition.

[2] P. N. Tan, M. Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education.

[3] Michael Berry and Gordon Linoff, "Data Mining Techniques", 2nd Edition Wiley Publications

[4] https://en.wikipedia.org/wiki/K-means_clustering

[5] https://en.wikipedia.org/wiki/Cluster_analysis