VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

LAB - 7 Exercises

Course Code	**:**	**CSE6017 – Mining Massive Dataset**	**Date**	**:**	**24/02/2020**
Lab Experiment	**:**	**K-Means clustering algorithm implementation in Apache Spark framework**	**Slots**	**:**	**L55+L56**
Instructors	**:**	**Prof. Ramesh Ragala**			

Objective:

1. To understand the K-Means clustering Implementation in Apache Spark using PySpark on Colab Environment

Exercises:

1. Implement the K-means clustering in iris dataset. Please do the following tasks while implementing the K-Means on Iris Dataset
   (A). Plot a graph on the raw data
   (B). use SQLContext for easy implementation
   (C). Store the iris dataset in Colab files
   (D). Identify and draw a plot to specify the optimal K value using elbow method.
   (E). Visualize the results after performing the k-means clustering with predicted value.
   (F). Provide the confusion matrix and accuracy of the result