SESSION PLAN	
Session Name	Machine Learning: Clustering/ k-means

Learning Outcomes(10 min)

- Differentiate between supervised and unsupervised methods
- Know the different types of unsupervised methods
- Understand how K-means and hierarchical clustering works
- Solve unsupervised problems using clustering

Prerequisites for the Mentor	**Prerequisites for the Student**
● Machine Learning: Clustering/ k-means - Go through the concepts in the platform.	● Machine Learning: Clustering/ k-means - Go through the concept and solve the tasks and assessments.

Timing	**Instructor Activities**

300 min	● Ask learners what they have learned from the concept? (10 min)
	● Overview of Machine Learning: Clustering/ k-means(60 min)
	  • K-Means clustering
	  • Hierarchical Clustering
	● Points of learning to reinforce through the session (when solving exercises and doing overview)
	● Key topic onboarding in this concept
	  ○ Spend a little bit of time before the session (5-10 mins) to engage the learners in a discussion on the difference between supervised and unsupervised learning. What would be major challenges in unsupervised learning as compared to supervised learning? (Nudge them towards the problem of evaluation in unsupervised learning). What would be the major advantage over supervised learning? (Nudge them towards the fact that the costly process of labelling can be avoided).
	  ○ Onboard the concept of clustering and bring out the differences from clustering and classification. A major point of confusion is the difference between clusters and classes w.r.t data.
	  ○ Definitely spend enough time taking the learners through the k-means algorithm and explain the mathematics behind the k-means algorithm. Currently, we are NOT taking the EM algorithm route and just onboarding the k-means without going into EM workflow.
	  ○ To make the learners understand the k-means intuition better, you can use the following visualization and show how the clusters change and finally converge. https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
	  ○ Go through the practical considerations of k-means and how to set the number of clusters.
	  ○ Talk about hierarchical clustering and a little about how the agglomerative clustering works. Spend some time talking about the different linkages to combine the clusters.
	  ○ At places where the coding exercises are present, it is recommended to solve the learning tasks / explain the approach to solve the learning tasks to help understand the learners how they are applying the theoretical concepts in code.
	  ○ Do not solve the assessment in session. Let the learners do the assessment themselves (in case they haven't done it).
	● Blog on Clustering:(7 min) https://medium.com/data-science-group-iitr/clustering-described-63e62833099e
	● Hierarchical clustering intuition:https://www.youtube.com/watch?v=0jPGHniVVNc (6.11 min)
	● Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After the first iteration clusters, C1, C2, C3 has the following observations: (10 min)
	  ○ C1: {(3,3), (5,5), (7,7)}
	  ○ C2: {(0,6), (6,0)}
	  ○ C3: {(6,6), (10,10)}
	  What will be the cluster centroids if you want to proceed for the second iteration?
	  (Solution: Finding centroid for data points in cluster C1 = ((3+5+7)/3, (3+5+7)/3) = (5, 5)
	  Finding centroid for data points in cluster C2 = ((0+6)/2, (6+0)/2) = (3, 3)
	  Finding centroid for data points in cluster C3 = ((6+10)/2, (6+10)/2) = (8, 8))
	● Practice problem on Machine Learning: Clustering/ k-means

	• Refer the GitHub repo for problems (30 min) ● Quiz on Machine Learning: Clustering/ k-means. (10 min) ● Questions and Discussion on doubts - AMA (30 min)

Context setting for code along (objectives and key takeaways) (5 min)

- Applying skills to solve a problem
    - Quiz learners on how to solve the problem posed given the concept that they have already learned. Let them come up with the approach.
    - Which data structure is best suited to capture data and calculate the result? Pose to the learner these questions.
- Adapting to something new
    - Bring attention to the learner about different formats of storing data and how to quickly search and implement how to read files stored in an unknown format to the learner.
    - How to look for help in documentation and quickly solve problems.
- Problem-solving workflow
    - Refer to Polya's How to Solve it - the broad principles of problem-solving.
    - Highlight how a hard problem can be broken down into smaller problems and the solution of the smaller problems build up as a solution to the larger problem

Code Along (120 minutes)

- Dataset overview - Credit Card(Customer Segmentation)
- The problem is a customer segmentation of credit card analysis.
- Spend some time establishing the business context and ask the learners what would be the business impact of solving the problem. Guide them to arrive at the ideal number of clusters they would want to arrive at and what are the desirable properties of the clusters of a credit card. If you were to do targeted marketing which segment of the customers would you target.
- The objective of this code along is to look at the various techniques of clustering and let them see for themselves the various clusters that are formed by the algorithms and their properties.
- High-level objective - what will be the outcome
- Explain the problem statement
- Engage the learner while solving the problem
    - While solving the problem pause, and question the learners if there are alternate ways of solving the problem.
    - While writing out the code, ask how to figure out in which data structure format is the data stored - use type()
    - Ask them which part of the data needs to be accessed to answer the questions posed in the code along.
- In case you fumble/are unable to get to the right answer - refer to the provided solution. Tell learners that it is ok to get stuck and how to look for help on StackOverflow, google
    - Purposefully make mistakes and ask the learners to point out the error and debug for you. Let them point out and

build the basic idea for the solution.

- Ask focused questions to gauge if learners are understanding
- Set the expectation that errors are important of the learning process and emphasis on the importance of debugging.
- Note questions parked if any. Resolve or answer later in slack or in the coming session

Next Session	

- Concept - Challenges in Machine Learning(30 min)
- Key topics to be highlighted - highlight where they would need to spend more time and importance w.r.t Data Science.
    - Different error metrics
    - Dealing with Imbalanced data
    - Dealing with small datasets
    - Values of K in K-Fold Validation
    - Optimal classifier choice