Instructions

1. You can solve Question-1 in a group. If you are interested to work as an individual, please do so. However, we encourage you to work as a group. Please refer to the "CUTe02-Batch37_team_allocation.xlsx" file to know your team members, ignore this, if you want to work as an individual. Each student must submit your/your team answers in the Grader tool on or before **Feb 17th 12:00 PM**. Your submission should contain R code and a PPT

2. Quesiton2, is an individual activity. Submit your **handwritten answers** during viva to your mentor. You have marks for neat handwriting. Submitting the answers after post viva session will not be accepted. Soft copies are not allowed.

3. 15 minutes Viva will be individual evaluation. (5 min for PPT and 10 min Q&A). Please refer to the "CUTe02-Batch37_team_allocation.xlsx" for your viva slot

4. Refer to the "CUTe02-Batch37_Evaluation_Metrics.xlsx" file for the evaluation criterion

5. Maximum marks: 100

1. Write the R code to solve the following

   a. You need to use the data set you prepared in the previous CUTe exam. Since there is a change in your team members, take the data set among your new team members that got the best MAPE.
   b. Understand the summary of the data set
   c. Split the data into Train and Test
   d. Build the regression model that gave you the Best MAPE and show the results on test data set
   e. Build the following models. Ensure you do the necessary data pre-processing before build the models
      i. Random forest
      ii. CART
      iii. C5.0
      iv. Ada Boost
      v. XG Boost
      vi. Support Vector Machines
   f. Apply these models on the test data set
   g. Please follow the steps below to compute the Error metric:
      i. Compute the median RMSE for each model
      ii. Choose the model that gave you min of median RMSE on the test data set

2. Answer any 5 Questions (one question from each of any 5 blocks) in your individual PPT, please prepare to answer any of these questions during Viva.

Block 1: Machine Learning

i. What are the techniques you will implement to improve the classification Accuracy?
ii. What is class imbalance problem and list out 5 methods to fix
iii. Explain the concept of Gradient descent and its importance in Machine learning?
iv. What is Bias and Variance concept in Machine learning and list out the approaches to fix them
v. Explain
    a. K-Fold validation
    b. Knn Classification
vi. What are the parameters that you will tune for each of the algorithms?
    a. Random forest
    b. CART
    c. C5.0
    d. Ada Boost
    e. XG Boost
    f. Support Vector Machines

Block 2: Collaborative Filtering

vii. What are recommendation Engines and the types of the recommendation engines
viii. How are KNN and collaborative filtering related to each other?
ix. Does KNN work well for high dimensional datasets? Why or why not?

Block 3: Decision Trees

x. Why is tree pruning useful in Decision trees?
xi. What is one problem associated with using information gain in decision tree learning?
xii. How are numeric attributes handled in decision tree learning? How do you do splits? How do you select the best cut-point for a binary split?

Block 4: Ensemble Techniques

xiii. Why do you think ensemble methods would work better than individual classifiers?
xiv. List out 5 differences between Ada Boost and Gradient Boosting Machines
xv. Ensembled models are known to return high accuracy. When do you think ensembled models fail in returning high accuracy?
xvi. What is the difference between bagging and boosting? and between adaboost and GBMs?
xvii. What is the difference between boosting and stacking?
xviii. In stacking, how is the meta learner learned? What are the features and the dataset to learn the meta-learner?

Block 5: Clustering Techniques

xix. Illustrate the strength and weakness of K-means in comparison with K-medoids
xx. Which one is not robust to outliers? KMeans, KMedoids or hierachical clustering?

xxi. Which one is faster? KMeans, KMedoids or hierachical clustering?
xxii. Give one method of deciding K for K-Means.
xxiii. How do we assess/evaluate the clustering algorithms performance?
xxiv. Suppose that the data mining task is to cluster points (with (x,y)) representing location into there clusters, where the points are

A1 (2,10) A2(2,5) A3(8,4) B1(5,8), B2(7,5) B3(6,4), C1(1,2) C2(4,9)

The distance function is Euclidian distance. Suppose initially we assign A1, B1, C1 as the centre of each cluster respectively. Use the k-means algorithm to show only the three cluster centres after second round of iteration

xxv. What is the value distance measure? How is it computed?

Block 6: Support Vector Machines

xxvi. What is concept of margin in support vector machines? (SVM)
xxvii. Mention pros/cons of various classifiers like SVMs, KNN, ensemble methods, decision trees.
xxviii. Does SVM suffer from the local minima problem? How about K-Means?
xxix. Define a SVM in one sentence.
xxx. How do you use an SVM for multi-class classification?
xxxi. What is the kernel trick in SVMs?
xxxii. Write the optimization problem for SVM to handle linearly separable data with noise.
xxxiii. How is class imbalance handled in SVMs?

3. Individual Viva. Questions will be only from above 6 blocks