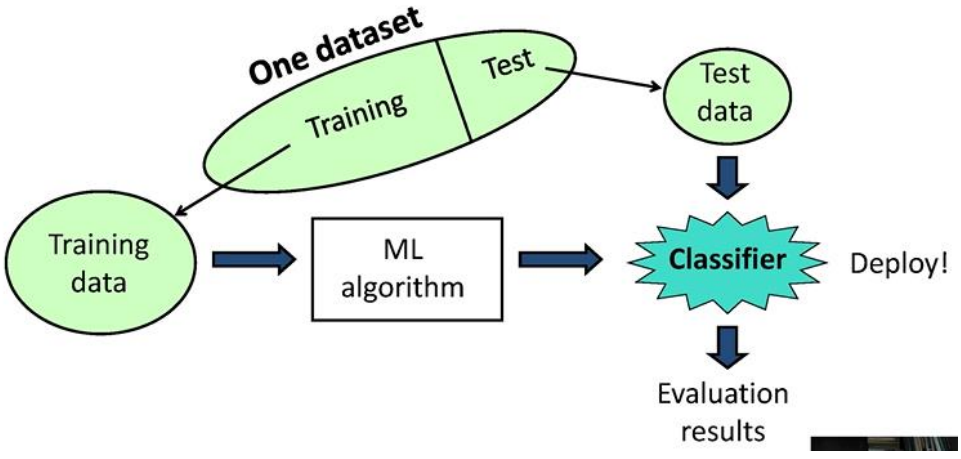
Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination(Online) - July 2020

Course: Data Mining **Semester: II**
Program: MBA(BA) **Time: 03 Hours**
Course code: DSBA 7008 **Max. Marks: 100**
Instructions: Use Weka to solve questions wherever required.

IMPORTANT INSTRUCTIONS

1. *The student must write his/her name and enrolment no. in the space designated above.*
2. *The questions have to be answered in this MS Word document.*
3. *After attempting the questions in this document, the student has to upload this MS Word document on Blackboard.*
4. ***Use Weka software wherever required and paste screen shot in document file for support of your answer***

		Marks	CO
Q1	**A) Download and open the anneal dataset.** i. How many *attributes* does it have? ii. Apply the unsupervised attribute filter *RemoveUseless* and find how many *attributes* does the dataset have now? iii. Identify one of the attributes that was removed by clicking *Undo* and then *Apply*. Now figure out why it was removed. B) **Explain the below terms along with the process of Machine learning as shown in diagram:** 	**10+10**	**CO2**
Q2.	**Classify the attribute 'Type' of glass.arff dataset using J48.** a) No. of instances and Attributes	**(2+2+5+ 5+6=20)**	**CO2**

	b) No of leaves and trees c) Write confusion matrix and explain it d) Write impact when Change minNumObj=15 e) Display Decision tree and interpret the model		
Q3.	**Open the *diabetes.arff* dataset and answer the following:** a) Select *Percentage split* as test option and set *percentage for training* to 80%. How many instances will be used for training, and how many for testing? b) Select the J48 classifier (default options) and evaluate it with the following seed values (*More options*): 1, 2, 3, 4, 5 c) What are the minimum and maximum values for the number of incorrectly classified instances? What is the mean of the accuracy for these five seed values? i. 63.3% ii. 75.3% iii. 76.0% iv. 95.0% d) What is the standard deviation of the accuracy for these five seed values? i. 2.1 ii. 2.8 iii. 3.1 iv. 3.3 e) If you did the experiment of (b) with 10 different random seeds rather than 5, how would you expect this to affect the mean and standard deviation? i. They would both stay about the same. ii. The mean would be a bit bigger but the standard deviation would be about the same. iii. The mean would be about the same and the standard deviation would be a little smaller. iv. Both the mean and standard deviation would be a bit smaller.	**5X4=20**	**CO2**
Q4.	**a) Open weather.normal.arff file.** i. Write step to remove 3rd attribute **ii.** Write step to remove High values of Humidity attribute **b) Open cpu.arff file.** i. Run Non- linear regression.	**2.5X2+3 X5 =20)**	**CO3**

	ii. Write all the linear equations with interpretation iii. Justify which one is better and why.		
Q5.	**Answer the following based on below given file:** @relation weather.symbolic @attribute outlook {sunny, overcast, rainy} @attribute temperature {hot, mild, cool} @attribute humidity {high, normal} @attribute windy {TRUE, FALSE} @attribute play {yes, no} @data sunny,hot,high,FALSE,no sunny,hot,high,TRUE,no overcast,hot,high,FALSE,yes rainy,mild,high,FALSE,yes rainy,cool,normal,FALSE,yes rainy,cool,normal,TRUE,no overcast,cool,normal,TRUE,yes sunny,mild,high,FALSE,no sunny,cool,normal,FALSE,yes rainy,mild,normal,FALSE,yes sunny,mild,normal,TRUE,yes overcast,mild,high,TRUE,yes overcast,hot,normal,FALSE,yes rainy,mild,high,TRUE,no a) Write the name of relation, attributes and number of instances. b) Name the class attribute and its labels. c) Write any two decisions list based on above data. d) Calculate number of possible instances. e) Draw decision tree based on above data.	**(3+ 2+ 5+ 3+ 7=20)**	**CO2**

Answers