DR. BABASAHEB AMBEDKAR TECHNOLOGICAL UNIVERSITY
LONERE – 402 103, RAIGAD (MS)
End Semester Examination, May 2018

Branch: M. Tech (CE / CSE / CS&IT / CS) **Semester: II**
Subject with Subject Code: Data Science [MTCE1201] **Marks: 60**
Date: 14 / 05 / 2018 **Time: 3 Hrs.**

==========

Instructions to the Students:

- Assume suitable data wherever necessary and mention it clearly
- If any of the parameter is noticed to be missing, assume appropriate value and mention it clearly
- Write answer to the point and correct.

Q. 1 Attempt the following.

a) Differentiate between Data Science, Machine Learning and AI using different criteria such as definition, scope and role. **(6 marks)**

b) What are various steps involved in an analytics of project? **(6 marks)**

Q. 2 Attempt the following,

a) What is text mining? Write R language code to convert uppercase to lowercase letters. **(6 marks)**

b) Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa. **(6 marks)**

Q. 3 Attempt the following.

a) What is categorical? Explain with suitable example the procedure to check the cumulative frequency distribution of any categorical variable. **(6 marks)**

b) What are the various aspects of a Machine Learning process? **(6 marks)**

Q. 4 Attempt the following.

a) What is K-means clustering? Enlist different steps in the K-means clustering algorithm. **(6 marks)**

b) A small accounting firm pays each of their five junior accountant's ₹35,000, two senior accountants ₹80,000 and the firms owner ₹3, 20,000 (all salaries are per month). **(6 marks)**

i) What is the mean salary paid per month?
ii) What is the median salary paid per month?
iii) How many employees earn less than the mean salary every month?

Q. 5 Attempt the following.

a) What is unsupervised machine learning? Enlist and explain steps in the unsupervised machine learning. **(4 marks)**

b) What is Linear Regression? **(4 marks)**

c) How machine learning is deployed in real world scenarios? **(4 marks)**

Q. 6

A leading fashion store chooses to predict the Willingness of a customer to buy a shirt of a particular price category based on the customers data. The company strongly believes that the willingness of a customer to buy depends on 3 factors essentially. They are the Gender of the customer (Male/ Female), the type of car used by the customer (Sports/ Family) and the type of shirt price category (Cheap/ Expensive).

From the past history, The Company has got the data of 12 customers (gender, The type of the car used by the customer and The Shirt price category) along with the data of whether they bought the shirt of that category. The company customer data is given below table. Then, answer the following questions.

Customer ID	**Gender**	**Car type**	**Shirt price category**	**Will buy?**
1	Male	Sports	Cheap	No
2	Male	Sports	Expensive	Yes
3	Male	Family	Cheap	Yes
4	Male	Family	Expensive	No
5	Male	Sports	Cheap	Yes
6	Male	Sports	Expensive	Yes
7	Male	Family	Cheap	Yes
8	Male	Family	Expensive	No
9	Female	Sports	Cheap	No
10	Female	Family	Cheap	No
11	Female	Sports	Expensive	No
12	Female	Family	Expensive	No

i) What is the best feature to split at the root level, If the splitting criterion is Entropy? **(4 marks)**

ii) Assume that you stop growing the tree, when there are 2 or fewer data points in a leaf node. If you use Entropy for the splitting criterion, Then How many leaf nodes will you end up with? **(4 marks)**

iii) Assume that you stop growing the tree, when there are 2 or fewer data points in a leaf node. If you use Entropy for the splitting criterion, then What is the prediction accuracy? **(4 marks)**

----- ##@@@@@## -----