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

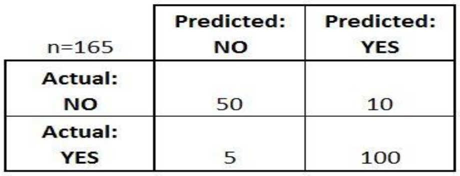
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, May 2021

Course: Machine Learning	**Semester: II**
Program: M. Tech - CSE	**Time 03 hrs.**
Course Code: CSAI7007P	**Max. Marks: 100**

SECTION A

1. Each Question will carry 5 Marks
2. Instruction: Multiple choice Questions.

<table>
<tr><th>S. No.</th><th></th><th>Marks</th><th>CO</th></tr>
<tr><td><b>Q 1</b></td><td>Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

<table>
<tr><th>n=165</th><th>Predicted: NO</th><th>Predicted: YES</th></tr>
<tr><td>Actual: NO</td><td>50</td><td>10</td></tr>
<tr><td>Actual: YES</td><td>5</td><td>100</td></tr>
</table>
Based on the above confusion matrix, choose which option(s) below will give you correct predictions?
1. Accuracy is ~0.91
2. Misclassification rate is ~ 0.91
3. False positive rate is ~0.95
4. True positive rate is ~0.95

A) 1 and 3
B) 2 and 4
C) 1 and 4
D) 2 and 3</td><td><b>5</b></td><td><b>CO1</b></td></tr>
<tr><td><b>Q2</b></td><td>Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.
Note: All other hyper parameters are same and other factors are not affected.
1. Depth 4 will have high bias and low variance
2. Depth 4 will have low bias and low variance

A) Only 1
B) Only 2
C) Both 1 and 2
D) None of the above</td><td><b>5</b></td><td><b>CO2</b></td></tr>
</table>

Q3	Which of the following options is/are true for K-fold cross-validation? 1. Increase in K will result in higher time required to cross validate the result. 2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K. 3. If K=N, then it is called Leave one out cross validation, where N is the number of observations. A) 1 and 2 B) 2 and 3 C) 1 and 3 D) 1,2 and 3	**5**	**CO3**
Q4	Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case? 1. Accuracy metric is not a good idea for imbalanced class problems. 2. Accuracy metric is a good idea for imbalanced class problems. 3. Precision and recall metrics are good for imbalanced class problems. 4. Precision and recall metrics aren't good for imbalanced class problems. A) 1 and 3 B) 1 and 4 C) 2 and 3 D) 2 and 4	**5**	**CO4**
Q5	For which of the following hyperparameters, higher value is better for decision tree algorithm? 1. Number of samples used for split 2. Depth of tree 3. Samples for leaf A)1 and 2 B) 2 and 3 C) 1 and 3 D) 1, 2 and 3 E) Can't say	**5**	**CO1**
Q6	Which of the following options can be used to get global minima in k-Means Algorithm? 1. Try to run algorithm for different centroid initialization 2. Adjust number of iterations 3. Find out the optimal number of clusters A) 2 and 3 B) 1 and 3 C) 1 and 2 D) All of above	**5**	**CO2**

<table>
<tr><td colspan="4"><b>SECTION B</b>
<b>1. Each question will carry 10 marks</b>
<b>2. Instruction: Write short / brief notes.</b></td></tr>
<tr><td><b>Q 7</b></td><td>a) Why is Nave Bayes classifier so powerful for text classification?
b) Why Normalization is required in machine learning?</td><td><b>6+4</b></td><td><b>CO1</b></td></tr>
<tr><td><b>Q8</b></td><td>a) In which algorithm, Ginni index is used. Explain the algorithm in detail with suitable example.
b) Why does the decision tree suffer often with overfitting problem?</td><td><b>6+4</b></td><td><b>CO3</b></td></tr>
<tr><td><b>Q9</b></td><td>a) What is the goal of SVM? How to select the margin?
b) Given the following data for the sales (in million dollars) of Car of an Automobile Company for 6 consecutive years.
<table>
<tr><td>Year</td><td>2013</td><td>2014</td><td>2015</td><td>2016</td><td>2017</td><td>2018</td></tr>
<tr><td>Sales</td><td>110</td><td>100</td><td>250</td><td>275</td><td>230</td><td>300</td></tr>
</table>
Based on the above data, predict the sales for next three consecutive years.

<b>OR</b>
A data set is given to you about utilities fraud detection. You have built a classifier model and achieved a performance score of 98.5%. Is this a good model? If yes, justify. If not, what can you do about it?</td><td><b>6+4</b></td><td><b>CO3</b></td></tr>
<tr><td><b>Q10</b></td><td>a) Which algorithm can be used to fit the data over a linear line? Is that algorithm supervised or unsupervised? And how would you calculate the cost for that algorithm?
b) Which is more important to you- model accuracy or model performance? Support with suitable example.</td><td><b>6+4</b></td><td><b>CO2</b></td></tr>
<tr><td><b>Q11</b></td><td>a) How could you divide the ‘training Set’ and ‘test Set’ in a Machine Learning Model? How much data will you allocate for training, validation, and test Sets?
b) Explain why k-fold cross validation does not work well with time series model? What can you do about it?</td><td><b>6+4</b></td><td><b>CO2</b></td></tr>
<tr><td colspan="4"><b>SECTION-C</b>
<b>Note: Attempt any one question from two options.</b></td></tr>
<tr><td><b>Q12</b></td><td>Differentiate between

a) Supervised, unsupervised and reinforcement learning
b) Bagging and boosting.
c) Linear Regression and Logistic Regression
d) Overfitting and under fitting

<b>OR</b>

Consider a medical diagnosis problem in which there are two alternative hypotheses:</td><td><b>10+10</b></td><td><b>CO4</b></td></tr>
</table>

	(1) that the patient has a particular form of COVID19 (+) and (2) That the patient does not (-). A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this COVID19. Determine whether the patient has COVID19 or not using the MAP hypothesis.		