SRN | | | | | | | | | | | | |

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	**UE20CS931**

December 2021: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER II

UE20CS931 - MACHINE LEARNING - II

Time: 3 Hrs	Answer All Questions	Max Marks: 100

INSTRUCTIONS

- Answer all the questions.
- Section A should be handwritten in the answer script provided and signed at the end of the same.
- Section B and C are coding questions which have to be answered in the system and uploaded in Olympus Login.
- Smartly use Grid SearchCV as it might impact the system performance.
- Write appropriate inferences.

Section A (30 marks)

1	a)	Define Precision and Recall.	5
	b)	Differentiate between bagging and boosting.	5
	c)	Describe ROC curve.	5
	d)	Explain Gini Impurity & Entropy.	5
	e)	Explain the working of Random Forest and list two advantages.	5
	f)	Explain Ensemble Learning	5

Section B (30 marks)

2		DATA DESCRIPTION: This data set contains details of a bank's customers who have taken loans and the target variable is a binary variable reflecting the fact whether the customer will payback the loan or not. CustomerId: Unique Ids for bank customer identification CreditScore: Credit score of the customer Geography:The country from which the customer belongs	

SRN | | | | | | | | | | | |

		Gender:Male or Female Age:Age of the customer Tenure: Number of years for which the customer has taken the loan Balance: Bank balance of the customer NumOfProducts: Number of bank products the customer is utilizing HasCrCard: Binary Flag for whether the customer holds a credit card with the bank or not IsActiveMember: Binary Flag for whether the customer is an active member with the bank or not EstimatedSalary: Estimated salary of the customer in Dollars Exited: Binary flag 1 if the customer closed account with bank and 0 if the customer is retained Target Column Description:'Loan' Payback the loan; class label is divided into groups (customer paybacks='1' and not='0').	
	(a)	Read the dataset and summarize important observations from the data set (i)Find out number of rows; no. & types of variables (continuous, categorical etc.) (ii)Calculate five-point summary for numerical variables (ii)Summarize observations for categorical variables – no. of categories, % observations in each category.	5
	(b)	Check for defects in the data. Perform necessary actions to 'fix' these defects. (i) Do variables have missing/null values? Is it a defect? If Yes, what steps are being taken to rectify the problem. (ii) Do variables have outliers? Is it a defect? If Yes, what steps are being taken to rectify the problem. (iii) Is the Target distributed evenly? Is it a defect? If Yes, what steps are being taken to rectify the problem. (iv) Is there any textual data? Is it a defect? If Yes, what steps are being taken to rectify the problem	10
	(c)	Summarize relationships among variables (i)Plot relevant categorical plots. Find out which variables are most correlated or appear to be in causation with Target? Do you want to exclude some variables from the model based on this analysis? What other actions will you take? (ii)Plot all independent variables with the target & find out the relationship? Perform the Relevant Tests to find out if the Independent variables are associated with the Target Variable. Hint: based on your observations you may want to transform features or create additional features.	5

SRN | | | | | | | | | | | | |

	(d)	Split dataset into train and test (70:30)	5
	(e)	Fit a base model and explain the reason of selecting that model. Please write your key observations. (i)What is the overall Accuracy? Please comment on whether it is good or not. (ii)What is Precision, Recall and F1 Score and what will be the optimization objective keeping in mind the problem statement. (iii)Which variables are significant? (iv)What is Cohen's Kappa Value and what inference do you make from the model?	5
		Section C (40 marks)	
3	(a)	How do you improve the accuracy of the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model.	20
	(b)	Summarize as follows (i)Summarize the overall fit of the model and list down the measures to prove that it is a good model. (ii)Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the relationship. Feel free to use charts or graphs to explain. (iii) What changes from the base model had the most effect on model performance? (iv) What are the key risks to your results and interpretation?	20