SRN □□□□□□□□□□□□

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

April 2022: 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

1. Answer all the questions.
2. Section A should be handwritten in the answer script provided.
3. Section B and C are coding questions which have to be answered in the system and uploaded in Google classroom.
4. Smartly use Grid SearchCV as it might impact the system performance.
5. Write appropriate inferences.

Section A (30 marks)

1	a)	Explain Decision tree Algorithm with an example.	5
	b)	What is confusion matrix? Mention all the evaluation metrics, which are computed from confusion matrix along with their formulas.	5
	c)	State all the assumptions associated with Logistic regression algorithm and Naive-Bayes algorithm.	5
	d)	Describe the advantages and disadvantages of Random forest, Ada boost and Gradient boosting algorithm.	5
	e)	Explain the concepts of bias and variance in Machine learning algorithms. Give some examples	5
	f)	Explain Supervised Learning.	5

Section B (30 Marks)

2

A non-profit healt organization has collected data about patients in a diabetic speciality hospital, recently has turned into a covid-cure center. The organization has to see if there is any pattern to classify the patients who has **survived** and **passed away**. Given below is the detail of the dataset.

Detail of the dataset
Patient I.D. --> Index of patient enrolment number
Age --> Age of the patient
Sex --> Gender of the patient
cp --> Chest Pain type chest pain type
trtbps --> resting blood pressure (in mm Hg)
chol --> cholestoral in mg/dl fetched via BMI sensor
fbs --> fasting blood sugar , which has been categorised into three group based on a different magnitude

		restecg --> resting electrocardiographic results thalachh --> maximum heart rate achieved Addiction --> Addiction of the patient keratin_type --> Keratin type of the patient Diabetes_type --> type of diabetes of the patient Hemoglobin_level --> level of hemoglobin of the patient blood_group --> Blood group of the patient Immunity_level --> Immunity level of the patient,which has been categorised into different group Affected_portion --> Portion of organ effect by virus Breath_ratio --> state air to the time of inhalation ratio. Survive --> whether a patient has overcome or passed away	
	(i)	Read the dataset and print the following * Shape of the data (1 mark) * Number of numerical and categorical variable (1 mark) * Descriptive stats of numerical data (1 mark) * Descriptive stats of categorical data (1 mark) * Summarize observations for categorical variables – no. of categories, % observations in each category. (2 mark)	6
	(ii)	Examine outliers by plotting. Examine is the Target variable evenly balanced	5
	(iii)	Perform appropriate encoding on the categorical attributes	5
	(iv)	Check for defects in the data like missing values, removing unnecessary features/columns. Perform necessary actions to 'fix' these defects.	6
	(v)	Examine the correlation and summarize the relationship between variables. Use appropriate plots to justify the same.	6
	(vi)	Split dataset into train and test (70:30)	2
		Section C (40 marks)	
3	(i)	Fit a base model and explain the reason of selecting that model. Please write your key observations. Compute F1 score, ROC score for the base model.	15
	(ii)	How do you improve the accuracy of the base model? Write clearly the changes that you will make before re-fitting the model. Fit the final model.	15
	(iii)	Summarize as follows with respect to the business problem 1.With respect to features (Which features are more important) 2.Evaluation metrics performances 3.Overall Results and Observations	10