SRN □□□□□□□□□□□□

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	**UE20CS905**

MARCH 2021: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER I

UE20CS905 - MACHINE LEARNING - I

Time: 3 Hrs	Answer All Questions	Max Marks: 80

1	a)	Explain Heteroscedasticity and Multicollinearity in Linear Regression.	2
	b)	Below is the equation of the births given by women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (bwght), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (cigs). The following simple regression was estimated using data on n = 1,388 births: **pred_bwght = 119.77 - 0.514*cigs** What is the predicted birth weight when cigs = 0? What about when cigs = 20 (one pack per day)? Comment on the difference.	2
	c)	Using the data in GPA2 on 4,137 college students, the following equation was estimated by OLS: **colgpa =1.392 - 0.0135*hsperc + 0.00148*sat** where colgpa is measured on a four-point scale, hsperc is the percentile in the high school graduating class (defined so that, for example, hsperc = 5 means the top 5% of the class), and sat is the combined math and verbal scores on the student achievement test. ● Why does it make sense for the coefficient on hsperc to be negative? ● Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large? ● Holding hsperc fixed, what difference in SAT scores leads to a predicted colgpa difference of .50, or one-half of a grade point? Comment on your answer.	3
	d)	How can you deal with autocorrelation of errors?	2
	e)	What is the difference between the Classification and Regression problem?	1

2	a)	● Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (roe, in percentage form), and return on the firm's stock (ros, in percentage form): ● **log(Salary) = 4.32 + 0.280*log(sales) + 0.0174*roe + 0.00024*ros.** The standard error of intercept is 0.32, sales is 0.035, se for roe is 0.0041 and se for ros is 0.00054. n =209 and R Squared = 0.283. The t-critical value at 10% is 1.282. ● By what percentage is salary predicted to increase if ros increases by 50 points? Does ros have a practically large effect on salary? ● Test the null hypothesis that ros has no effect on salary against the alternative that ros has a positive effect. Carry out the test at the 10% significance level. ● Would you include ros in a final model explaining CEO compensation in terms of firm performance? Explain.	6
	b)	If y = 2x1 + 12x2 + 3x3 + 5 is the linear regression equation, then explain how the coefficients of x2 and x3 affect the value of y.	2
	c)	Explain Gradient Descent in brief.	2
3	a)	DATA DESCRIPTION: The data set consists of complete educational details of students right from their schooling to MBA and previous work experience. Our main objective is to predict the Salary of the students based on the info available ATTRIBUTES: ● SlNo - ID of the student ● Gender - Gender of Student ● Percent_SSC - Percentage of marks scored in SSC ● Board_SSC - Types of Boards in SSC ● Percent_HSC - Percentage of marks scored in HSC ● Board_HSC - Types of Boards in HSC ● Stream_HSC - Specialization in HSC ● Percent Degree - Percentage of marks scored in Degree ● Course_Degree - Different courses in degree ● Experience_Yrs - Work Experience of the Students ● Entrance_Test - Test which students give for MBA college Entrance ● Percentile_ET - Percentage of marks scored in Entrance_Test ● Percent_MBA - Percentage of marks scored in MBA ● Specialization_MBA - Specialization in MBA ● Marks Communication - Percentage of marks scored in Communication ● Marks_Projectwork - Percentage of marks scored in Project Work ● Placement - Whether Student got placed or not	15

		● Salary - Salary of students Perform Exploratory data analysis and summarize important observations from the data set Some pointers which would help you, but don’t be limited by these i. What are the number of rows; no. & types of variables (continuous, categorical etc.) ii. Calculate five-point summary for numerical variables iii. Summarize observations for categorical variables – no. of categories, % observations in each category iv. Check for defects in the data. Perform necessary actions to ‘fix’ these defects. check for numerical/ categorical variable(wrong representation) missing values, outlier treatment, skewness, encoding v. Summarize relationships among variables. vi. Split dataset into train and test (70:30)	
	b)	Fit a base model. Please write your key observations i. What is the overall R2? Please comment on whether it is good or not. ii. What is the adjusted R2? Is it different from R2? Why? iii. Which variables are significant. Explain the Feature Selection Technique used? iv. Is there multicollinearity, Suggest ways to remove it. v. Which other key model output parameters do you want to look at?	15
4	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. Please feel free to have any number of iterations to get to the final answer. Marks are awarded based on the quality of the final model you are able to achieve.	25
	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?	5