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: 4 Hrs	Answer All Questions	Max Marks: 80

1	a)	State few applications of Machine Learning	2
	b)	What is the difference between Classification and Regression problem?	2
	c)	Mention any two assumptions of linear regression algorithm.	2
	d)	How can you deal with multicollinearity?	2
	e)	Explain some issues with Machine Learning.	2
2	a)	Explain Forward Selection in brief.	2
	b)	If $y = 2x_1 + 12x_2 + 3x_3 + 5$ is the linear regression equation, then explain how the coefficients of x_2 and x_3 affect the value of y.	2
	c)	How can you handle overfitting and underfitting?	2
	d)	Explain Gradient Descent in brief.	2
	e)	What is Lasso Regularization?	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	15

- 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
- 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

a. What are the number of rows; no. & types of variables (continuous, categorical etc.)

b. Calculate five point summary for numerical variables

c. Summarize observations for categorical variables – no. of categories, % observations in each category

d. 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

e. Summarize relationships among variables.

f. Split dataset into train and test (70:30)

b)	Fit a base model. Please write your key observations	15

a. What is the overall R2? Please comment on whether it is good or not.

b. What is the adjusted R2? Is it different from R2? Why?

c. Which variables are significant?

d. Is there multicollinearity?

e. Which other key model output parameters do you want to look at?

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.	25

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.

	b)	Summarize as follows	5

a. Summarize the overall fit of the model and list down the measures to prove that it is a good model
b.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.

c.What changes from the base model had the most effect on model performance
d.What are the key risks to your results and interpretation?