SRN □□□□□□□□□□□□

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE20CS905

May 2022: 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

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 Olympus Login.
4. Smartly use GridSearchCV as it might impact the system performance.

		Section A (20 marks)	
1	a)	Write a note on feature scaling in Data Science.	4
	b)	Explain assumptions of linear regression.	4
	c)	Explain Box-cox transformation.	4
	d)	Write a note on Backward elimination method of feature selection algorithms.	4
	e)	How can be the problem of overfitting can be reduced in Linear regression? Explain the different variants to solve the over fitting problem in Linear regression.	4
		Section B (30 Marks)	
2		Problem Statement: A firm is trying to predict the temperature based on various environmental and seasonal settings. They want to utilize rational approach facilitated by machine learning in predicting the temperature in Celsius. They collected a data set with different features along with the temperature. The challenge is to learn a relationship between the important features and the temperature and use it to predict the future temperature. Develop a machine learning model to predict the temperature using the features provided in the dataset. Prior to building the ML model EDA need to carried out to understand and clean the data.	
	(i)	Read the dataset and perform the following *Read the dataset. (1 marks) *Observe the data types of the features. (1 marks) *Observe the features in the dataset that add little to no information. (1 marks) *Visualize the relationship between 'Wind Speed (km/h)'and 'Temperature (C)'. (1 marks)	4
	(ii)	Perform the following EDA for understanding the data set * Check for Missing values in the data. (1 marks)	4

		* Display the percentage of missing values in each column. (2 marks) * Implement a strategy to deal with the missing values. (1 marks)	
	(iii)	Create a new column as year, month, day and hour using the date column.	4
	(iv)	Perform the below tasks. * Use boxplot to visualize the outliers of numeric columns. (3 Marks) * Perform outlier elimination using IQR method. (3 Marks)	6
	(v)	Perform the below tasks * Plot a correlation plot and highlight the correlations through color map. (3 Marks) * Compute multicollinearity for input features using VIF and drop the high VIF features sequentially. (3 Marks)	6
	(vi)	Perform the below tasks * Drop all irrelevant features though the EDA performed in above steps. (3 Marks) * Use encoding technique to encode the categorical variables. (3 Marks)	6
Section C (30 marks)			
3	(i)	Perform the below tasks. *Split the data into 2 parts train and test (70:30) (2 Marks) *Use OLS stats models package to build the Linear Regression model (2 Marks) *Generate the summary report. (2 Marks)	6
	(ii)	Build a model using sklearn's least squares regression. Interpret the coefficients (4 + 2 Marks)	6
	(iii)	Find the best set of significant variables from the dataset using forward selection technique, backward elimination technique. Also, display the R-squared score for the model built using forward selection technique selected variables. (2 + 2 + 2 Marks)	6
	(iv)	Validate models performance using 5 fold cross validation and print the different RMSE scores. Comment about model's overfitting. (3 + 3 Marks)	6
	(v)	Use Grid Search CV to determine the optimal value of alpha if Lasso regression is used to build the model. (3 + 3 Marks)	6