SRN												

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

August 2023: 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: 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 Olympus Login.
4. Smartly use GridSearchCV as it might impact the system performance.

		Section A (20 marks)	
1	a)	What is Multicollinearity? How to detect the presence of multicollinearity and which variables are involved in it?	4
	b)	Explain the procedure involved in k-fold cross validation.	4
	c)	Explain the assumptiona of linear regression.	4
	d)	Explain the procedure involved in Forward Feature Selection.	4
	e)	How the problem of overfitting can be reduced in Linear regression? What is bias variance trade off?	4
		Section B (40 Marks)	
2		Problem Statement: Housing price dataset of Bengaluru city is provided. Based on the given details predict the price of the house. Below are features details - area_type: The type of the house area feature 'total_sqft' specifies. - availability: The availability date or availability status of the property. - location: The locality of the property in Bengaluru city. - size: The size of the housing property in BHK (or Bedrooms etc.,). - society: The name of the Apartment. This name is encrypted for confidentiality. - total_sqft: The 'area_type' area of the property. - bath: Number of bathrooms available in the house. - balcony: Number of balcony/balconies the house has. - price: Price of the housing property in Lakhs. (target feature) .	
	(i)	Read the dataset and perform the following 1. Read/load the dataset as a pandas Dataframe.(1 mark) 2. Print/show the dimensions of Dataframe i.e., no of rows and columns. (1 mark) 3. Print/show the data types of all the features/columns. (1 mark) 4. Print/show statistical summary of all the numeric featurs. (1 mark)	8

		5. Print/show statistical summary for all the categorical variable. (2 marks) 6. Find out Feature wise Missing value counts. (2 marks)	
	(ii)	Perform Below Exploratory Data Analysis(EDA) Tasks. 1. Show/Visualize the relationship between fetures 'bath' and 'price' using scattered plot. (1 marks) 2. Show/Visualize the relationship between fetures 'balcony'and 'price' using scattered plot. (1 mark) 3. show/Visualize the relationship between fetures 'bath','balcony' and 'price' using 3D Scatterplot. (2 marks) 4. Show outliers distribution of variable 'bath' by drawing Boxplot. (3marks)	7
	(iii)	4. Pre-process the Dataframe as Mentioned Below. (25 marks) 1. Replace missing values of the feature 'balcony' with numerical value 0 and convert its feature type to int.(2 marks) 2. Replace missing values of the feature 'bath' missing values with numerical 1 and convert feature type to int.(2 marks) 3. Replace missing values of the feature 'location' with a constant "missing".(2 marks) 4. Replace missing values of the feature 'society' with a constant "missing".(2 marks) 5. Convert the feature 'size' to int by removing alphabetic content and keep only numeric content. In case of missing/null content replace by constant numeric value- 2. (3 marks) 6. Convert the feature 'total_sqft' to numerical using 'to_numeric' method. Also, replace all its missing entries by mean.(3 marks) 7. Eliminate all the outlies records/rows from Dataframe with respect to feature'bath' (2 marks) 8. convert 3 categorical features i.e. 'availability', 'location' and 'society' into numerical using label encoding. (6 marks) 9. Perform one hot encoding on feature 'area_type' , also ensure output columns are of type int (3 marks)	25
		Section C (40 marks)	
3	(i)	Perform Below Modeling Tasks (15 marks) 1. Split the processed Dataframe into 2 parts train and test with ratio as 70:30. Ensure feature 'price' as target(y). (3 marks) 2. Use OLS statsmodels package to build the Linear Regression model on the train set. Also,generate the summary report. (6 marks) 3. Using sklearn's linear regression model train model on the train set and interpret the coefficients. (6 marks)	15
	(ii)	Model Comparisons and Hyperparameter tuning 1. Train below models and obtain values using 5 fold cross validation on train data and 'RMSE' metric. Find the metric (RMSE) score in test set and suggest the best model. (15 marks) - Ridge (alpha = 1, max_iter = 500) (5 marks) - Lasso (alpha = 0.01, max_iter = 500) (5 marks) - ElasticNet(alpha = 0.1, l1_ratio = 0.01, max_iter = 500) (5 marks) 2. Using Random search on Lasso model find the best value of alpha and corresponding RMSE value on test set. (10 marks)	25