|  |  |  |  |
| --- | --- | --- | --- |
|  | **PES University, Bengaluru**  (Established under Karnataka Act No. 16 of 2013) | | **UE20CS905** |
| **November 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.   1. Smartly use GridSearchCV as it might impact the system performance. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Section A (20 marks)** | | | |
| 1 | a) | How the problem of overfitting can be reduced in Linear regression? What is bias variance trade off? | 4 |
| b) | Write about the different variance measures involved in the operation of Linear Regression. | 4 |
| c) | Explain the assumptions of linear regression. | 4 |
| d) | Discuss the use of Rsquared and adjusted Rsquared in a Linear Regression Model.? | 4 |
|  | e) | What is k-fold cross validation? Write briefly about the procedure. | 4 |
|  | | | |
| **Section B (40 Marks)** | | | |
| 2 |  | Problem StatementPredicting the cost of claims in an insurance company is a real-life problem that needs to be solved in a more accurate and automated way. Several factors determine the cost of claims based on health factors like BMI, age, smoker, health conditions and others. So here the problem statement is to predict the insurance claims by the policyholders. Data Definition  **age** : Age of the policyholder (Numeric)  **sex:** Gender of policyholder (Categoric)  **weight:** Weight of the policyholder (Numeric)  **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight (Numeric)  **no\_of\_dependents:** Number of dependent persons on the policyholder (Numeric)  **smoker:** Indicates policyholder is a smoker or a non-smoker (non-smoker=0;smoker=1) (Categoric)  **claim:** The amount claimed by the policyholder (Numeric)  **bloodpressure:** Bloodpressure reading of policyholder (Numeric)  **diabetes:** Indicates policyholder suffers from diabetes or not (non-diabetic=0; diabetic=1) (Categoric)  **regular\_ex:** A policyholder regularly excercises or not (no-excercise=0; excercise=1) (Categoric)  **job\_title:** Job profile of the policyholder (Categoric)  **city:** The city in which the policyholder resides (Categoric)  **hereditary\_diseases:** A policyholder suffering from a hereditary diseases or not (Categoric) |  |
| (i) | **Basic Pandas DataFrame Operations to understand the data**  1. Read/load the dataset. (1 mark)  2. Print the dimensions/shape of the dataset i.e. no of rows and columns.(1 mark)  3. Print the data types of all the features. (1 mark)  4. Does any feature need type casting? if yes, perform the same and change the datatype to the correct type.(2 marks)  5. Print statistical summary of all the numeric as well as categorical features. (1 mark)  6. Find out how many unique categories are available and number of instances in each of the categorical features (2 mark) | 8 |
| (ii) | **Perform Below Exploratory Data Analysis(EDA) Tasks.**  1. Check/Visualize the relationship between the numerical features with the target feature 'claim'. State the inferences (3 marks)  2. Check/Visualize the relationships between categorical features and the Insurance claim. (3 marks)  3. Do you spot any multicollinearity among the numerical Independent features? (1 mark) | 7 |
|  |  |  |
| (iii) | **Perform Data pre-processing as Mentioned Below. (25 marks)**  1. Check and visualize if there is any missing values feature wise. (2 mark)  2. Impute the missing values of the feature 'age' by group wise mean across the gender column. (2 marks)  3. Replace missing values of the feature 'bmi' by mean imputation. What issues can be faced when mean imputation is directly used to a feature.(3 marks)  4. Handle the 0 (zero) values present in 'bloodpressure' column. (as bloodpressure cannot be 0) (2 marks)  5. Check and visualize if there are any outliers present in the numerical features. (2 mark)  6. Treat the outlies records/rows from Dataframe with respect to the concerned features. (8 marks)  (a) Use IQR and Drop the outliers in the feature 'bloodpressure'.  (b) Use appropriate transformation function to treat the outliers in 'claim' feature.  7. Convert the categorical features into numerical using the following encoding techniques.( 6 marks)  (a). Encode 'sex' column using replace function  (b). Encode 'hereditary\_diseases' using pandas get\_dummies() function.  (c). Encode 'job\_title' using target encoding (without using category encoders library) | 25 |
| **Section C (40 marks)** | | | |
| 3 | (i) | **Perform Below Modeling Tasks (15 marks)**  1. Split the pre-processed data frame into 2 parts train and test with ratio as 80:20. Ensure feature 'claim' as target(y). (3 marks)  2. Use OLS statsmodels package to build the Linear Regression model on the train set.Generate the summary report. Write your inferences (6 marks)  3. Predict the values using test set (3 marks)  4. Compute accuracy measures RMSE and MAPE (3 marks) | 15 |
| (ii) | **Model Comparisons and Hyperparameter tuning (25 marks)**  1. Using sklearn's linear regression model train model on the train set and Interpret the coefficients. (5 marks)  2. For the above model compute bias error (mean rmse) and variance error across 5 fold cross validation (5 marks)  3. Train below models and obtain values using 5 fold cross validation on train data and 'rmse' metric. Find the metric score in test set and suggest the best model. ( 10 marks)  - Ridge(alpha = 1, max\_iter = 500) (5 marks)  - Lasso(alpha = 0.01, max\_iter = 500) (5 marks)  4. Using Grid serach on Lasso model find the best value of alpha and corresponding rmse value on test set. ( 5 marks) | 25 |