School of Computer Science Engineering and Technology

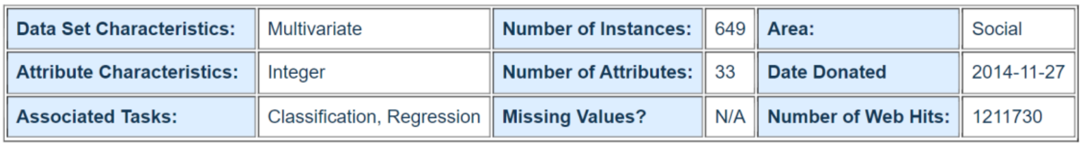
Course- BTech	Type- Core
Course Code- CSET301	Course Name-AIML
Year- 2022	Semester- Even
Date- 05-09-2022	Batch- V Sem

3 - Lab Assignment No. (3.2_2)

Exp. No.	**Name**	**CO-1**	**CO-2**	**CO-3**
3.2_2	Multiple Linear regression	✓	✓	--

Objective: Build Multiple Linear Regression Model using Sklearn

1. About Dataset: This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modelled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details). Problem Statement is to predict student performance in secondary education (high school). (5)



Data Set Characteristics:	Multivariate	Number of Instances:	649	Area:	Social
Attribute Characteristics:	Integer	Number of Attributes:	33	Date Donated	2014-11-27
Associated Tasks:	Classification, Regression	Missing Values?	N/A	Number of Web Hits:	1211730

2. Download the dataset 'Student Performance' provided by UCI Machine Learning repository.
   Dataset link: https://archive.ics.uci.edu/ml/datasets/student+performance(5)
3. Read the data and store the features in X and output variable in Y. Consider the following features only in X from the downloaded dataset:(10)
   Features (X)
   1) age - student's age (numeric: from 15 to 22)
   2) address - student's home address type (binary: 'U' - urban or 'R' - rural)
   3) famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
   4) reason - reason to choose this school (nominal: close to 'home', school 'reputation','course' preference or 'other')
   5) studytime - weekly study time (numeric: 1 -<2 hours, 2 - 2 to 5 hours, 3 - 5

to 10 hours,or 4 - >10 hours)

6) failures - number of past class failures (numeric: n if 1<=n<3, else 4)
7) schoolsup - extra educational support (binary: yes or no)
8) famsup - family educational support (binary: yes or no)
9) paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes orno)
10) activities - extra-curricular activities (binary: yes or no)
11) higher - wants to take higher education (binary: yes or no)
12) internet - Internet access at home (binary: yes or no)
13) romantic - with a romantic relationship (binary: yes or no)
14) freetime - free time after school (numeric: from 1 - very low to 5 - very high)
15) goout - going out with friends (numeric: from 1 - very low to 5 - very high)
16) health - current health status (numeric: from 1 - very bad to 5 - very good)
17) absences - number of school absences (numeric: from 0 to 93)
18) G1 - first year math grades (numeric: from 0 to 100)
19) G2 - second year math grades (numeric: from 0 to 100)

Output target (Y)

20) G3 - final year math grades (numeric: from 0 to 100, output target)

4. **Data Pre-processing step: (40)**
   a) Check the presence of missing values. If there is any missing value present use a suitable approach to remove them.
   b) Check the presence of Categorical columns. Handle it if present. i.e., Transform categorical features into numerical features. (Hint: Use eitherone hot encoding, label encoding or any other suitable pre-processing technique).
   c) Split the dataset into train and test (Hint: Use train_test_split class, Use Splitting Ratio: 70:30, 80:20 Criterion)
   d) Scale the numerical columns value using minmax_scale() or any other scaling function.
5. Train Multiple Linear Regression Model on training dataset (Hint: sklearn.linear_model.LinearRegression class) (20)
6. Check the performance of model using 'Mean Squared Error' (MSE) and R2-Score (Hint: sklearn.metrics Lib)(10)

Suggested Platform: Python: Jupyter Notebook/Azure Notebook/Google Colab.