Mini Project - Regression

- ❖ Starting Date : 10th September 2023
- ❖ Last Date : 17th September 2023
- ❖ Weightage : 24 Marks
- ❖ Submissions : File upload in Canvass

Insurance Cost Prediction

Problem Statement: Predict the Insurance Cost based on given dataset

Dataset The dataset contains 1338 rows of insured data, where the

Insurance charges are given against the following attributes of the insured:

1. Age
2. Sex
3. BMI
4. Number of Children
5. Smoker
6. Region

The insurance company now wants to predict the cost for the new customer with the help of above mentioned attributes. As a result, they want you to build a prediction model which can correctly set the insurance cost of the new client provided the attributes are given. The task involves the following things:

TASKS & Weightage:

- ❖ **Analyze the dataset and do EDA with proper interpretation (Exploratory Data Analysis) – 4 Marks**
  - Target variable distribution
  - describe() and info() of the dataframe
- ❖ **Perform the various plotting techniques to identify the correlation relationship with proper observation – 2 Marks**
  - pairplot (available in Seaborn library)
  - Correlation b/w all features (See corr() function of dataframe) and heatmap of this correlation matrix
- ❖ **Split the dataset into training and test sets. 1 Marks**

  Case 1 : Train = 80 % Test = 20%

  [x_train1,y_train1] = 80% ; [x_test1,y_test1] = 20% ;

Case 2 : Train = 10 % Test = 90%

[x_train2,y_train2] = 10% ; [x_test2,y_test2] = 90% ;

- Refer train_test_split from sklearn

❖ **Perform the model Building using Multiple Linear Regression for case 1 and case 2 –12 Marks**

[Case 1 : Train = 80 % Test = 20% Case 2 : Train = 10 % Test = 90%]

- Normal SK-Learn library – 3 Marks
- Gradient Descent – 3 Marks
- Stochastic Gradient Descent – 3 Marks
- Mini Batch – 3 Marks

❖ **Calculating the performance metrics for each model with respect to each case - 2 Marks**

- Track R2 score and RMSE for each case, each model

❖ **Compare the overall result and write the observations for each model and provide conclusion – 3 Marks**

NOTE: EDA refers to exploring the dataset from various facets such as outliers, correlations, wrong data types, Null values etc.

Some useful links are given below for your reference. You can refer to them while writing your own code.

Link 1 - https://towardsdatascience.com/gradient-descent-in-python-a0d07285742f (Links to an external site.)

Link 2 - https://medium.com/coinmonks/implementation-of-gradient-descent-in-python a43f160ec521 (Links to an external site.)

(Links to an external site.)

Link 3 - https://www.geeksforgeeks.org/ml-mini-batch-gradient-descent-with-python/ (Links to an external site.)