

Spectrum Club,
College of Engineering and Technology,
Bhubaneswar
www.spectrumcet.com

Dear Intern,

Welcome to **Data Science and Machine Learning Task 2**. We are finally at the stage where we have the fundamentals required to pre-process and dataset and make predictions on it. In this Task we will be building a student grade prediction based on the student's data and previous marks. This will be the final project of your internship and the evaluation shall be based on how much accuracy you can get on your model.

Technology Stack to be used:

In this final task, we will mainly be working on encoding nominal types of data (as given in the student.txt file), converting them to numerical form, and then creating a machine learning model to predict the score of a student given his attributes as well as to optimize the model. The libraries we will mainly be using will be :-

- sklearn
- numpy
- Pandas
- statsmodel.api

Final stage:






This stage is divided into 2 parts. Part one will be encoding the categorical values of the dataset, creating x as input features and y as output column. Part 2 will be creating a multiple regression machine learning model on it, and optimizing the model.

Tasks:

Billy has his data refined. But he sees there are some words which he can't feed the computer. He wants you to help him clean the dataset so that he can finally feed the computer the values and predict his final exam marks on it. Help him to do.

1. For all the nominal values given in the dataset, like Pstatus, reason and every other column which has text values, convert all of them into numeric values by using either sklearn.preprocessing's library "LabelEncoder" or "OneHotEncoder". (Interns are expected to go through each of the methods and apply whichever they feel preferable. The evaluation won't be affected by the type of method used.)

- 
2. Create a column called “final_grade” in the same way as you did in the 2nd task. (**Note: Don’t delete the columns G1 G2 and G3 as we did in the 2nd task**)

Our model shall take all the input features of a student and should be able to predict the final grade of the student accurately. For that we need to create input and output features to feed into the model.

So,

3. Initialize a variable y which shall contain the output column, i.e., final_grade, as an array(you can use numpy for this).
4. Initialize a variable x which shall contain every other column except “G3”.
5. Using sklearn.model_selection library’s “train_test_split”, split the dataset into 4 parts, x_train, x_test, y_train and y_test.

Refer to the 2nd doc provided for the 2nd part of this task.





RESOURCES: -

The resources provided here include both documentation and Youtube tutorials for the above tasks. If any case of any confusion, do Google once and search around or ask in the forum, but do understand the concepts behind the functions properly. Remember, Stackoverflow is your best friend for programming.

- **ML using Sklearn –
(Documentation)**

<https://scikit-learn.org/stable/>



(Video Tutorial)

<https://www.youtube.com/playlist?list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9r>

<https://www.youtube.com/watch?v=hJ2sKPj5Xn4>

(Videos 1-16 of the first link should be enough to get you the basic idea of Machine Learning and how ML pipelines are created. The 2nd link is about the label encoding and one hot encoding methods.)



The dataset has been provided with this zip file, and the attributes, and the type of data they contain(numeric, non-numeric/binary, nominal), etc. have been provided too.

Good luck!

LAST DATE OF SUBMISSION: 31th May-2020

WARM REGARDS,
SPECTRUM, CET-B