Exam for Machine Learning Python Lab

Consider the file provided with the assignment and execute the analysis described below according to the best practices of Machine Learning. You are allowed to use *only the computers of the lab, use the operating system Ubuntu*, you are not allowed to use any other device, email or any other messaging tool. You can use *only the websites accessible through the computers of the lab, as listed in the following page.*
Cooperative work will be heavily sanctioned

The notebook must operate as follows:

1. Load the file `pollution.csv`, explore the data showing size and synthetic information . . . . **1 pt**
2. keep only the columns

   `'Date','Time','CO(GT)','NOx(GT)',`
   `'PT08.S5(O3)','NMHC(GT)','NO2(GT)','C6H6(GT)'`

   then drop all the rows containing null values and substitute the columns `'Date','Time'` with a new column `'DateTime'` obtained encoding the two strings ; you can do this using the `to_datetime` method of pandas with an appropriate format string; use as target `'C6H6(GT)'` and use for the prediction the other five numeric columns; the `'DateTime'`; the date will not be used for the prediction, but only for the presentation of the results, as required below . . . . **2 pt**
3. fit a linear regression model on the training data, evaluate RMSE and R–square on the training and on the test data, show the values . . **3 pt**
4. plot the true and the predicted target values for the test set, using for the x axis the `'Date'` column . . . . **2 pt**
5. fit a polynomial regression model trying several maximum degrees, up to 4, optimising the model with cross validation evaluate RMSE and R–square on the training and on the test data, show the optimal degree of the polynomial and the results . . . . **4 pt**
6. plot the true and the predicted target values for the test set, using for the x axis the `'Date'` column . . . . **1 pt**
7. comment the results of the two models . . . . **3 pt**

Total points for tasks 16

Quality of the code **4pt**

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalised
- Remove the code you use for testing and inspecting the variables during the development
- Naming style of variables must be uniform and in English
- Bad indentation and messy code will be penalised
- Non generalised solution, such as three sequential statements with the same kind of operation instead of a loop, will be penalised

Total grade:20
Additional directions, the assignments not compliant with the rules below will not be considered:

- The notebook name must be
  `yourworkplace_youremailusername.ipynb` in lowercase letters
  E.G. if your worplace is `lab9_35` and your email is `mario.rossi45@studio.unibo.it`, the notebook filename will be `lab9_35_mario.rossi45.ipynb`
- The solution must directly access the data in the same folder of the notebook, the name of the file must be the same as the file provided.
- Upload the notebook only to `http://eol.unibo.it` in the activity specified by the teacher, any other way of submitting the notebook will be ignored

Allowed websites

- `https://numpy.org`
- `https://scipy.org`
- `https://pandas.pydata.org`
- `https://matplotlib.org`
- `https://seaborn.pydata.org`
- `https://scikit-learn.org/stable`