CST4050 - Component 2

General information

This coursework is about developing a supervised learning model on the dataset released on UniHub.
Your submission comprises of a *individual report*.
You are required to submit your work via the dedicated Unihub assignment link by the specified deadline.

Note that this link will '*timeout*' at the submission deadline. *Your work will not be accepted as an email attachment if you miss this deadline*. Therefore, you are strongly advised to allow plenty of time to upload your work prior to the deadline.

Individual report

The World Health Organization (WHO) keeps track of the health status for many countries in the world. The purpose of this challenge is to create an explainable supervised learning model to predict the outcome variable `Life_Expectancy` from other health factors collected in 193 countries of the world by WHO in 2015. You need to use an appropriate programming language (such as Python) to implement your designed pipeline.

Your individual submission needs to contain the following sections:

1. Loading data and preliminary analysis
2. Training and testing a machine learning pipeline
3. Tuning the proposed machine learning pipeline (*Optional*)
4. Model interpretation (*Optional*)
5. Discussion: pros/cons and time complexity of the designed pipeline

Students are required produce a Notebook containing, on the same document, (*i*) your detailed comments, (*ii*) your program, (*iii*) its corresponding output, and (*iv*) your output interpretation.

Students are required to use the appropriate *template* available on the same UniHub folder (provided in `.ipynb` format) and to export it to *PDF*. You cannot change the provided template and you need to use exactly the same sections as provided in your template, otherwise you will get 5 marks penalty and any new section not included in your template will be ignored.

When submitting your individual report, please adhere to the following rules:

- Fully motivate every block of code of your Notebook.
- Comment each result you get from your Notebook.
- Export your Notebook file into a *PDF* file. Upload only your *PDF* file to UniHub.

Check your marking scheme to understand how your individual report will be evaluated.

Marking Scheme

Your submission will be evaluated according to:

- *Loading data and preliminary analysis*: 15%
  Target and features well introduced and discussed.

- *Training and testing a machine learning pipeline*: 20%
  Supervised machine learning model correctly trained and tested.
  Good motivations provided.
  Sound solution proposed.
  Variance-bias trade-off discussed.

- *Tuning the proposed machine learning pipeline*: 10%
  Good motivations provided.
  Correct use of k-fold cross validation.
  Correct use of the validation set.
  High quality comments on the bias-variance trade-off.

- *Model interpretation*: 10%
  Good procedure proposed.
  Good explanations.
  Correct and sound interpretation of results.

- *Discussion: pros/cons and time complexity of the designed pipeline*: 15%
  Time complexity of the proposed model discussed.
  Time complexity of the whole proposed pipeline discussed.
  Critical analysis on the pros and cons of the proposed pipeline.

- *Tidiness and clarity*: 20%
  Clear and easily understandable comments.
  Good amount of comments provided, clarifying what each block of code does.
  Well discussed results from each block of code.
  Short and sharp submission.

- *Technical proficiency in English*: 10%
  High quality submission, free from spelling, grammar and punctuation errors.
  Ensure that you have carefully proofread your work prior to submission.