**Adarga Data Scientist NLP Task – Document Classification**

**Task:**

Download the dataset attached, the original source from Kaggle is also linked below. It contains excerpts from school textbooks across seven different topics. Your task is to create a machine learning based classifier model to predict which topic each document belongs to as accurately as possible. You don’t need to work with all seven topics, use only the Maths, History and Computer Science related documents and your predictions should be across these three classes (if you prefer to work with all seven instead that’s okay as well, just indicate it).

Your solution should be in Python, in the form of either a pure python script or a Jupyter notebook. You may use any third-party libraries, frameworks, tools to work out your solution, deep learning and classical approaches are both accepted, showing your thinking is more important than the solution you choose and how well it works. You may also consult blogs, stack overflow, etc to help with your solution but if you copy-paste any piece of code please indicate it with a comment pointing to the source.

Some of the steps and considerations to think about are:

* Data cleaning and preparation, why is it important?
* Model choice, what made you choose a particular method?
* Evaluation, which metric is appropriate to use?
* How could results be improved?

Please leave brief comments throughout your work, the easier we understand what you did the more you shine as a candidate. You will be asked questions about your solutions at the interview.

**Data Source:** <https://www.kaggle.com/datasets/shivamkushwaha/bbc-full-text-document-classification>