**ADVANCE DATA SCIENCE PROJECT (TTS)**

**Project title:** TBD

**Members: Esther**, George, Kofi, and Me

**Aim/Objectives**

This project aims to create a multivariate machine learning (ML) predictive model that predicts two or more dependent variables (DV, i.e., the Y’s) at the same time from the same set of independent variables (IV, i.e., the X’s). The DVs must have at least one (1) categorical and one (1) continuous variable. The IVs can be a mixture of categorical and continuous variables.

We hope to investigate different techniques to facilitate creating a model that will classify and regress over a set of IVs. The motive is to create a model capable of estimating both the type (class) and size (magnitude) of a criterion we hope to investigate.

**Scope**

1. Find a multivariate dataset from a legitimate repository
2. Perform an exploratory data analysis on it using pandas or other python packages, for example, k-lib.
3. Preprocess the data and perform feature engineering (FE) to investigate the influence of different techniques on our model.
4. Select multiple ML algorithms capable of active combine classification and regression.
5. Create a pipeline to build the model combining the ML algorithms, feature engineering techniques, and tunning of the algorithms hyper-parameters.
6. Optimize the model by permutating over the different FE techniques and the hyper-parameters to obtain the optimum model(s).
7. Perform Y-scrambling to test the robustness of the model.
8. Save the model using pickle for future use.
9. If time permits, we can create a Streamlit app from this project.

**Candidate datasets**

1. Wisconsin Breast Cancer Prognosis (Classification **and/or** Regression)
2. Wine Quality (Classification **or** Regression)
3. Students’ Performance (Classification **and/or** Regression)

**NOTE:**

We don’t have to achieve everything on the list for the scope section. This write-up is just a suggestion.