# DATA 201 – Assignment 3

Total marks: **20**

Due date: **11:59pm, Friday 23 September**.

Submit **code** and **outputs** in a **single PDF file**.

**Background:**

Financial inclusion remains one of the main obstacles to economic and human development in developing countries. For example, in East Africa, only 14% of adults have access to or use a commercial bank account. Access to bank accounts has been regarded as an indicator of financial inclusion. It is also an essential contributor to long-term economic growth.

You are requested to develop a machine learning model using a survey dataset to predict which individuals are most likely to have or use a bank account. You will have to predict the ***probability*** of having or using a bank account for each individual.

The model will be developed using file **train.csv** and tested on file **test.csv**. The description about the data is given in the table below.

| Variable | Description |
|---|---|
| uniqueid | Unique identifier for each interviewee |
| bank_account | If interviewee has access to a bank account |
| location_type | Type of location |
| cellphone_access | If interviewee has access to a cellphone |
| household_size | Number of people living in the house |
| age_of_respondent | The age of the interviewee |
| gender_of_respondent | Gender of interviewee |
| relationship_with_head | The interviewee’s relationship with the head of the house |
| marital_status | The martial status of the interviewee |
| education_level | Highest level of education of the interviewee |
| job_type | Type of job interviewee |

**Requirements:**

- Use area under the ROC (AUC) as the evaluation metric. [**2 marks**]
- Load the dataset, determine the target column, remove irrelevant variables (if any) and explore the training set to gain insights. [**3 marks**]
- Try at least 3 different machine learning models (e.g., add pre-processing transformers, use their default hyperparameters, etc.), and estimate the performance of the models on unseen data (e.g., using cross-validation). [**3 marks**]

- Select one model, optimise it (e.g., add/remove features and/or redesign the pipeline if you wish, perform hyper-parameter tuning, etc.), and (re-)estimate the performance of the model. [**7 marks**]
- Test the final model on the test set, report the AUC-ROC and at least 4 other evaluation metrics (e.g., AUC-PR, accuracy, sensitivity, specificity, etc.) [**4 marks**]
- Include a discussion section at the end of your Notebook (about what you have learnt, difficulties, what have worked and not worked, future directions, etc.). [**1 mark**]

**Notes:**

- Write **your name and student ID** at the beginning of your notebook.
- After completing your work, use menu item ***Kernel => Restart & Run All*** in Jupyter, then print your notebook including code and outputs to a **PDF file** for submission. Chrome may be the best browser for this. If you are not happy with the presentation format of the PDF file, you may want to use menu *File => Download as => HTML (.html)* then print file .html to PDF.
- You can use any public Python package.
- The requirements above have no order that you have to follow.
- In order to compute AUC-ROC and AUC-PR, your model needs to make predictions in probability or score (i.e., a number in [0, 1], for example, by using function `predict_proba()`). However, to compute accuracy and other metrics, you also need to predict classes (e.g., using `predict()` or thresholding the predicted probabilities).
- Use your own assumption and judgement if you are unsure about any information in the dataset. However, remember to mention it in discussion.
- Try to write functions for all data transformations you apply, try feature engineering (e.g., creating new features), and try to automate all the steps as much as possible (e.g., using custom transformers, etc.). You may have **bonus mark**s for this; however, your total mark will not excess **20**.