PPOL670-Introduction to Data Science
PROJECT PROPOSAL
MARIA ARNAL CANUDO
TOPIC: FINANCIAL INCLUSION IN SUB-SAHARAN AFRICAN COUNTRIES

Rationale

This project aims to investigate the variables that determine the financial inclusion of adult population in Sub-Saharan African countries. In order to perform a comprehensive analysis of the factors, the analysis looks at a wide range of variables related with several topics, from education to health, job market or macroeconomics variables.

The reason of including several indicators from different topics relies on the complexity of the topic to analyze. Financial inclusion is not an isolated factor determined by a specific element or context, but it is a complex status that relies on several elements. Thus, these variables vary from one country or context to another. However, making an analysis of several similar countries in terms of economic level and society structure we would be able to compare and make a conclusion about the most important categories that determine the financial inclusion. Hence the success of this project depends on finding those elements that are determinants or more relevant for the financial inclusion in Sub-Saharan African countries.

Data

Data from the World Bank (WB) and International Monetary Fund (IMF) will be used to perform this project. Specifically, the Africa Development Indicators dataset (WB). This dataset provides indicators at country level on health, education, job market, utilities (access to electricity, fuel, internet, mobile money, etc.), natural resources, technology, agriculture, access to financial institutions, industry and services, among others. Even though there are macroeconomics variables as well, to complement the WB dataset, the

Financial Access Survey from the IMF[1] will provide the macroeconomic categories. Both datasets are at country level. However, they require wrangling: turn both datasets to wide or long, remove variables are useless for the analysis (i.e. population below 15 years, duplication of categories between two datasets, non-Sub-Saharan African countries, etc.) and merge them and manage the missing values among others. Besides, despite the differences in gender in lower income countries, the goal of this project is not to distinguish between women and men, but a previous analysis of the distribution of some variables among both will be performed in order to ensure a balance.

The analysis will be based on a range of 10 years (2008-2018). The selection of these years relies on the positive economic trends of several African countries during this period of time, which might indicate an increase of financial inclusion.

Empirical strategy

In addition to the wrangling data process, a machine learning statistical model on random forest will be performed to find those variables that predict better the outcome and are correlated among themselves. In addition, this system would allow us to perform an analysis by group of categories such as macroeconomic indicators, education, health and gender in case the previous data analysis is unbalanced. Thus, each tree would provide an average of the prediction.

Furthermore, graphs on the trends of the outcomes as well as distribution of the variables, correlations and other possible outcomes based on the results will be depicted to support the interpretation of the results.

[1] Financial Development might be interested to include, but it is an index. Please, advise.