**HUDK 5053 Feature Engineering Studio**

**- Deliverable 2: Problem Design Statement (September 22, 2016) –**

M.S. Learning Analytics

Sunmin Lee

Before joining TC, my working area focused on educational statistics, especially with developing countries. While working with ministry people who were in charge of national educational statistics, I found that we haven’t done much analysis of currently available data. At the same time, one of the big issues in education sector globally is to reduce out-of-school children (in here, I will narrow it to drop-out students). Hence, the problem I am addressing is that there are too many drop-out students in basic education (i.e. K-12) worldwide without knowing the reason and only a few studies have been done by countries to find out why. My educational goal to address this issue is to provide a better quality education to all children in K-12 and to reduce unnecessary cost and energy to bring children back to school.

Based on this, my objectives for this project will be to conduct a pilot study with available international public data (e.g. UNICEF, OECD) to develop a prediction model. My priority in this research is to hope that this pilot study can be a milestone in this area since not many studies are done with international data. Lastly, I am planning to measure who are the expected drop-outs (hopefully we can have estimated number since we cannot know a specific student) and what kind of characteristics, in other words, features they have.

Attached is a fake dataset referenced from UNICEF MICS (Multiple Indicator Cluster Survey) data. MICS is the “largest source of internationally comparable public data on women and children” (UNICEF MICS, 2016). It has been used as the main source for tracking Millennium Development Goals (MDGs) indicators and will continue for new 2030 Sustainable Development Goals (SDGs). Main features that I have in here is ‘where the student is living (Area)’, ‘whether his/her parents are educated’, ‘wealth’, ‘student’s literacy status’, ‘highest level of education attended’, ‘age at beginning of school year’, ‘he/she is engaged in activities for house income’, ‘he/she is engaged in household chores’, etc. Most of them are categorical questions and I created for 1,000 data points. With this possible features, I want to develop a prediction model to identify which students (if impossible, how many) are in danger of drop-outs and what is the main factors reasoning that.

* The problem you are addressing
* The educational goal that you plan to address
* Your objectives for the project
* Your priorities for the project
* What you plan to measure to achieve your educational goal and address your problem