Final Project Report

Determinants of Youth Employment. Sub-Saharan Africa

Introduction

Youth employment is a major concern for governments across the world, considering that young people are up to four times more likely than adults to be unemployed. In developing countries this concern is even stronger given that young people face additional risks such as conflict, black market, crime, drugs, etcetera. Several types of interventions around the world to increase youth employment have been applied, in both, either supply and demand side, however, the consensus is growing that more integrated approaches are needed.

One of the questions in the literature is which are the elements that allow young people to transit from unemployment to employment. In 2003, the International Labour Organization (ILO) developed the School-to-Work Transition Survey (SWTS) that provides some information to try to answer this question. Using this Survey, the objective of this project is to try to identify if there are particular characteristics that young people who get a job have with respect to those that do not have one. This can provide some policy elements for potential interventions to link young people with labor market.

Finally, it is important to say that this project analyze the results of the surveys applied to more than 27,300 young people in 9 countries in Sub-Saharan Africa.

Road map of the Project

First, additional background information for the project will be provided. Then, information about the data, how the databases were gotten and initial data wrangling manipulations will be explained.

As a third point, I will explain some additional wrangling I did after running machine learning techniques, such as removing outliers and missing values. In a fourth point I will show the results of a single logit regression before showing the results after applying machine learning techniques. In the penultimate part of this document, you will read the discussion around the best predictor model from the ML techniques and what those it means. Finally, I will discuss if I succeed in the terms established in the project proposal.

Problem Statement and Background

Youth employment is a major concern of governments across the developing world, where young people are up to four times more likely than adults to be unemployed. In these

countries, 85% of the jobs taken by youth belong to informal markets, and a quarter of the youth cannot find a salary above the extreme poverty line, $1.25 per day (Goldin et al., 2015). In addition, the majority of youth are dissatisfied with their employment situation, and want to change jobs, even more so in rural settings and in the agricultural sector (OECD, 2018). For all sectors, the main reason for wanting to change jobs is low pay, followed by the temporary nature of employment and poor working conditions.

The 2013 World Development Report (World Bank, 2018) comes to the conclusion that supply-side interventions are likely to be less effective and sustainable without addressing institutional and market failures on the demand-side that affect the entry and growth of firms. According to a 2018 World Bank study, "promoting job creation requires targeted interventions on the demand side that address specific constraints or market failures facing informal sector enterprises, formal sector firms, and farms. Important firm specific constraints or market failures include, insufficient access to finance, capacity and information gaps, coordination failures, and failure to capture social externalities of jobs."

To strength future interventions, it is relevant to know who the young are, which are their characteristics and in particular, which are the differences between those who have a job and those that do not. Those difference can be demographic or socioeconomic. This is exactly the aim of this project, try to see if from the SWTS those characteristics can be found.

Data

Source of the data

With the aim to portray which are the elements that allow a youth to transit from unemployment to employment, the International Labour Organization (ILO) designed the School-to-Work Transition Survey (SWTS). As ILO defined, the SWTS is a unique survey instrument that generates relevant labor market information on young people aged 15 to 29 years, including longitudinal information on transitions within the labor market.

The databases for 33 countries were shared to me by the ILO for this final project. The databases corresponds to countries from different continents, however, for this project, 9 countries from the non Sub-Saharan Africa were chosen: Benin, Liberia, Togo, Congo, Malawi, Zambia, Madagascar, Tanzania and Uganda.

The reason for choosing Sub-Saharan Africa is because I am currently working in my capstone project and it is related to youth employment in Burkina Faso. The countries selected preserve similar income conditions as Burkina. It is relevant to say that in the Project Proposal, I mentioned that I would work with the information from the 33 countries for which I have information, however, due to time constraints and because the databases are not exactly the same, I focused in the countries mentioned before.

The unit of observation in this project is each youth among 15 and 29 year old that answered the survey in each of the selected countries.

Variables of interest

In average, each survey has around 400 questions. Considering that my objective is to find which characteristics determine that a youth get a job, I chose as dependent variable the dummy "employed". This will allow me to structure models that measure the probability that a youth get a job based on the chosen independent variables.

As independent variables, I chose 13 variables: age, sex, education level, marital status, field of study, if worked while studying, if did internship, if had a reserve wage, if used to live in rural or urban areas, if have children, financial situation at home, and the level of education of both, father and mother. Unfortunately, after wrangling the data and running the ML techniques, I dropped educational level, reserve wage, internship, previous residence area and field of study. Finally, I dropped parents level of education due to inconsistencies in the way the bases are labeling.

Wrangling data

The most challenging part in wrangling the data was to match each variable of interest in the different databases. Even though each database follows in general the same SWTS structure and almost all the questions are the same, the names and labels are different. This implied to go over each database, along the more than 400 questions in the 9 countries, to match each variables of interest.

Each base was in Stata format, so I changed that to R format and applied a "remove labels" command because labels merged with tittles, complicating the subsequent merge. Then, for each database, I selected just the variables of interest mentioned above and relabeled those. Then I created dummies or categorical variables.

Finally, I merged all the 9 databases in one, getting a N size of 27,365, distributed this way: Benin, 4306; Congo, 3276; Liberia, 2416; Madagascar, 3300; Malawi, 3097; Tanzania, 1988; Togo, 2708; Uganda, 3049 and Zambia, 3225.

Extra wrangling

In the process of running ML techniques, I identified some outliers in the data and variables with many missing values.

Variables with outliers:Remove age>29. Remove financial_sit>5

Variables with many missing values:edu_level: 25,529; reserve_wage: 20,853; internship: 18,730; prev_resid: 21,518; field_edu: 16,182

Given this, I built a new database, removing the problematic variables. As a final manipulation, I removed all the missing values.

Analysis

The model

$$Y = \beta_0 + \beta_n X_n + \epsilon$$

Where:

- Y is the probability of a youth to being employed
- X_n is a matrix of predictors based on the information from the survey

At this point, we know that our predictors are: **Sex of the youth**: 1 if women, 0 if male. This coefficient will provide information about gender gap, this is, if there is a higher probability for a male to get a job. **Age**: this is a numeric variable between 15 and 29 years. We expect that the older the higher the probability to get a job. **Married**: 1 if married, 0 if not. We expect a higher probability to have a job for those who are married with respect to those who are not. **Work while study**: 1 if worked while studying, 0 if not. We expect that those that worked while studying have higher probability to be employed. **Having children**: 1 if they have kids, 0 if not. We expect that those who have kids have a higher probability to have a work. **Financial situation**: categorical variable that increase from the easiest situation to a tough one. We expect a positive relation, the hardest the situation, the strongest the motivation to get a job.

Logit regression

This is a simple logit regression to see the the results of the expected model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1009857	0.0994434	-31.183426	0.0000000
sex	0.0656467	0.0293698	2.235177	0.0254057
age	0.1009160	0.0040088	25.173315	0.0000000
married	0.2629989	0.0451201	5.828861	0.0000000
work_while.study	1.1907479	0.0365527	32.576192	0.0000000
child	0.7864839	0.0351944	22.346819	0.0000000
financial_sit	0.1077483	0.0146186	7.370633	0.0000000

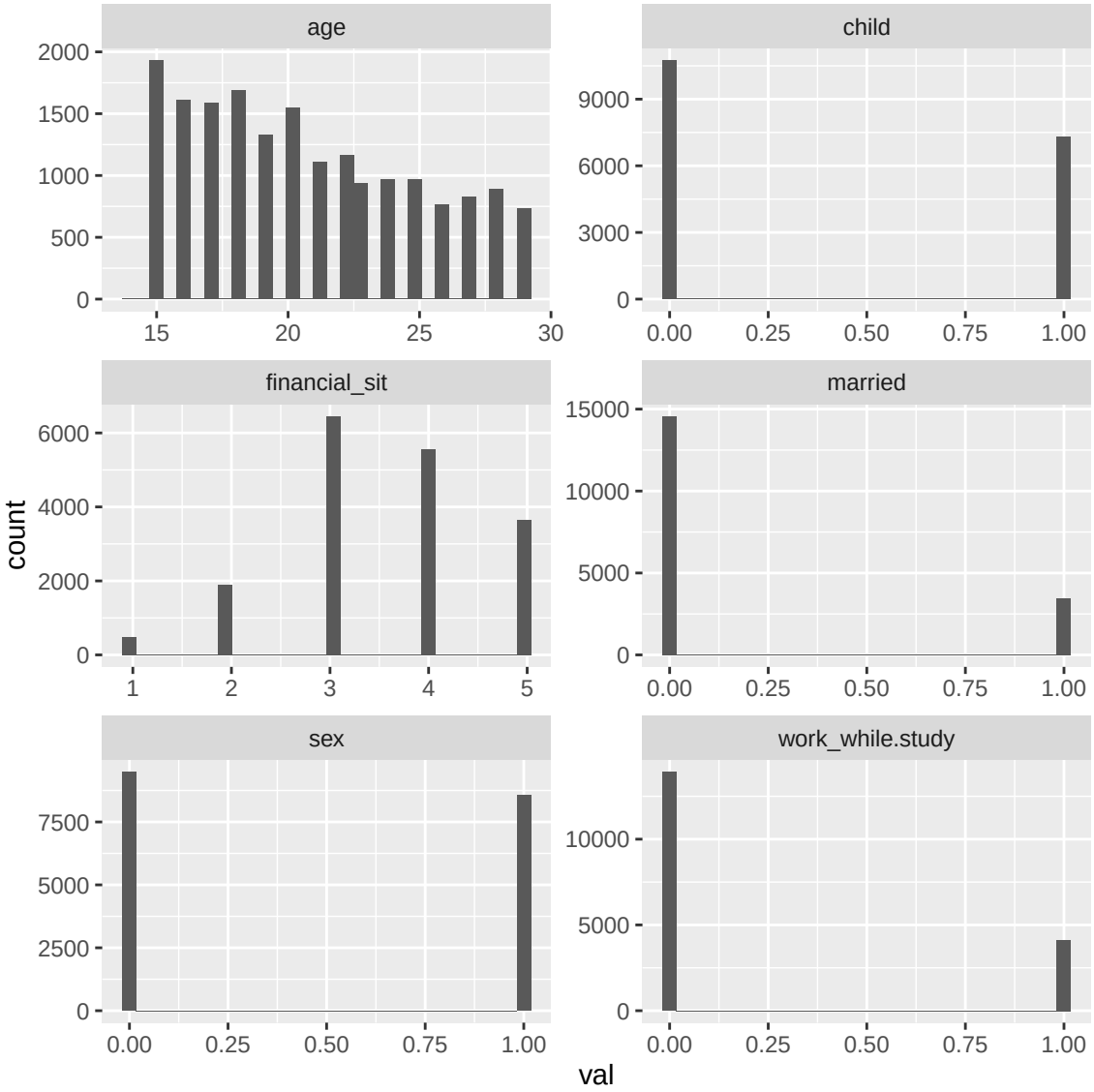
Looking at the coefficients from the logit regression, all of them threw the expected results excepting for sex. Recall that sex is labeled as 1 for female, therefore, this result says that there is a slightly higher probability to have a job, in average, if being a female. From the other results, it is interesting to see that those that worked while studying and those with children have strongly higher probabilities to have a job with respect to those that did not worked or have no children.

Machine learning Techniques (MLT)

As mentioned before, the MLT follow the same model but under supervised techniques. These are used to predict or explain the relation among variables. We will apply 4 techniques: a new logistic regression, K-Nearest Neighbors, Classification and Regression Trees and Random Forest. Along each technique an explanation will be provided.

The first step is to split the data in training (80%) and test (20%) data set. Since this is a random partition, it is expected that the results would be the same in both partitions.

After the partition, let's take a look on the distribution for each variable for the train data set. Most of them are binary so we expect to have data in zero and one. Doing this before, allowed me to identify outliers and missing values.



The changes to our data are saved in a "recipe" so we can then apply this instructions to other data sets. In this case we apply the "recipe" to our train and test data. Some of the changes that I did before was to change the dependent variable from numbers to text, that it

is from 1 to "yes" and 0 to "no". As well, although not relevant for this porpoise, the variable "country" that refers to the country of each youth, change from a categorical variable to a dummy for each country.

Cross-validation

When comparing different machine learning models, we want to be sure that we're making a fair and equal comparison. One way that random chance can sneak into our assessments of model fit is through cross-validation. We want to make sure that we're cross-validating the data on the exact same data partitions. In this case, I used the k-fold cross-validation method with 5 folds in the data, getting groups of around 3,600 observations.

```
## Fold1 Fold2 Fold3 Fold4 Fold5
##  3608  3609  3609  3609  3610
```

Then, I set up the validations conditions. Here we establish that all the process that will be run in the MLT, have to consider a cross validation among the 5 folds we just created. Likewise, I set up that we are dealing with a classification problem, that it is, that our outcome is either have a job or not.

Models

To run the models, I used a powerful R package named "caret". The performance measure that I will use to determine the model that best perform is the Receiver Operating Characteristics or ROC, given that this is a classification problem.

ROC is a probability curve. It tells how much a model is capable of distinguishing between classes. Higher the area under the curve, better the model is at predicting 0s as 0s and 1s as 1s. In our example, to identify if a unemployed youth is correctly classified as unemployed and if an employed one is correctly classified as employed.

Logistic Regression

The difference in this logistic regression from the first showed above is that here we are only focus in the ROC, in order to compare with the other models. The ROC for this model is 0.7416589.

```
## Generalized Linear Model
##
## 18045 samples
##     6 predictor
##     2 classes: 'no', 'yes'
```

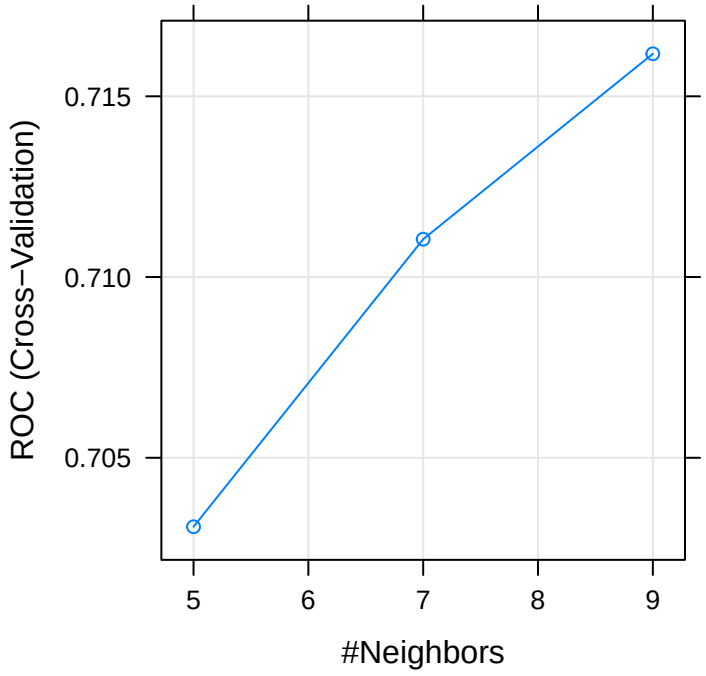
```
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3608, 3609, 3609, 3609, 3610
## Resampling results:
##
##   ROC        Sens       Spec
##   0.7399705  0.6905203  0.6740083
```

K-Nearest Neighbors

This algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other and are classified in "neighbors". This model plays with the number of K neighbors and provide a ROC for each of them.

As we will see in the results and in the graph below, a k=9 provides a higher ROC than k=5 but, once we go over k=10, the increase in ROC turns smaller for each additional neighbor. The highest ROC=0.7212098 for this model was with k=9, higher than in the logistic regression.

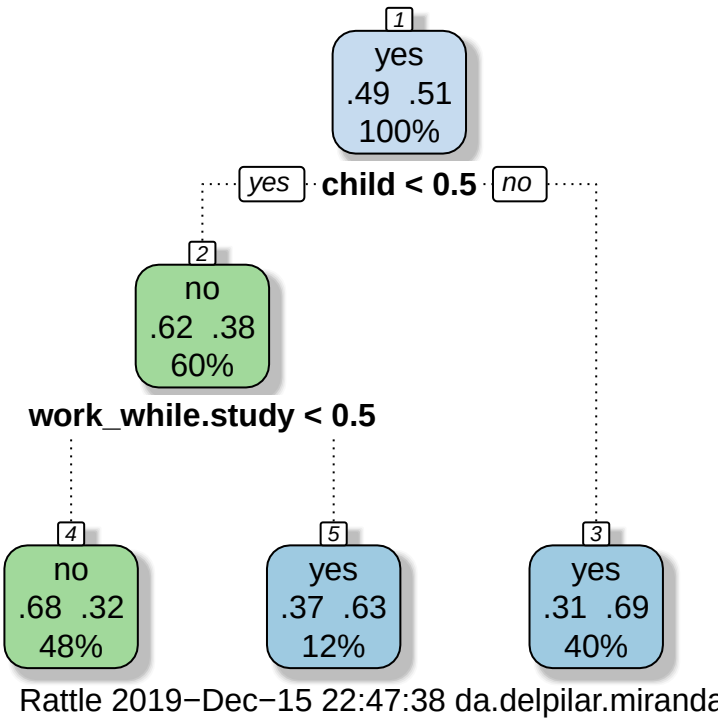
```
## k-Nearest Neighbors
##
## 18045 samples
##     6 predictor
##     2 classes: 'no', 'yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3608, 3609, 3609, 3609, 3610
## Resampling results across tuning parameters:
##
##   k  ROC        Sens       Spec
##   5  0.7030922  0.6683487  0.6517915
##   7  0.7110467  0.6645925  0.6621739
##   9  0.7161761  0.6661062  0.6692142
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```



Classification and Regression Trees (CART)

This model follows the logic of binary decision trees and the model is used to identify the "class" within which a target variable would likely fall into. In this example, the model identified that the class "having children" and "work while studying" were the best classifications to predict the probability of a youth to get a job.

Decision tree:




```
## n= 18045
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 18045 8919 yes (0.4942643 0.5057357)
##   2) child< 0.5 10749 4124 no (0.6163364 0.3836636)
##     4) work_while.study< 0.5 8599 2765 no (0.6784510 0.3215490) *
##     5) work_while.study>=0.5 2150  791 yes (0.3679070 0.6320930) *
##   3) child>=0.5 7296 2294 yes (0.3144189 0.6855811) *
```

Random Forest

A Random Forest process is basically running several CART models or decision trees, playing with different variables and different sets of the database, taking the average of all of them instead of just choosing one.

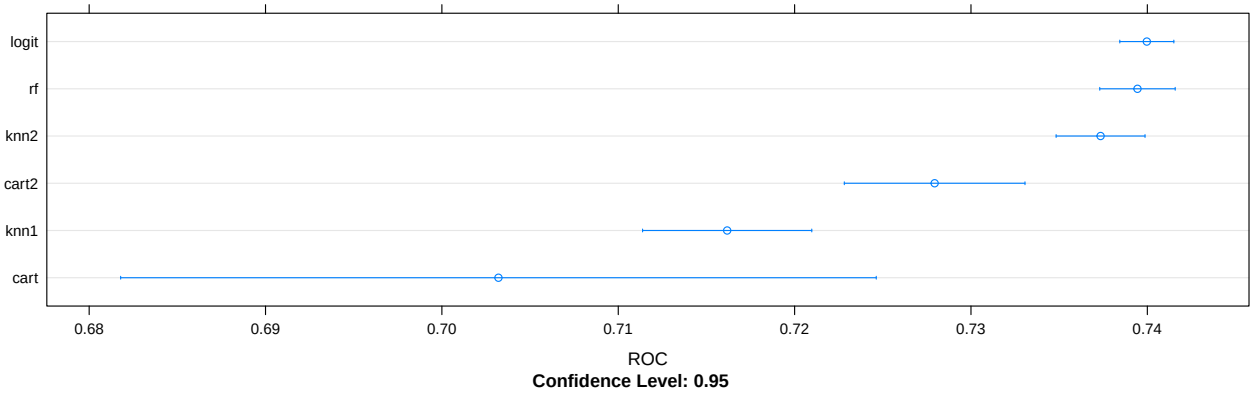
As we can see in the output table, the RF with the highest ROC is 0.7414514 with 2 predictors randomly selected and nodes splitted by "extratrees" as tuning parameter.

```
## Random Forest
##
## 18045 samples
##     6 predictor
```

```
##     2 classes: 'no', 'yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3608, 3609, 3609, 3609, 3610
## Resampling results across tuning parameters:
##
##   mtry  splitrule   ROC        Sens       Spec
##   2     gini        0.7377132  0.6749918  0.6861710
##   2     extratrees  0.7394407  0.6726931  0.6931018
##   4     gini        0.7053676  0.6760289  0.6419567
##   4     extratrees  0.7155873  0.6785519  0.6566401
##   6     gini        0.6868110  0.6817471  0.6171107
##   6     extratrees  0.6878443  0.6825880  0.6191924
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 2, splitrule =
##  extratrees and min.node.size = 1.
```

Results

Based on the next plot, I conclude that the MLT that provides the higher ROC, this is, the highest correct classification of employed and unemployed young population based on the selected independent variables are the Random Forest model and logit regression. The difference in ROC is slightly different in favor of logit model, however, I will stick with RF as my predilection model.



In the next confusion matrix we can see those young people that was correctly classified as employed or unemployed running the predict parameters in the Random Forest over the test data. The classification accuracy is 67%.

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   no  yes
##        no  1494  692
##        yes  735 1589
##
##                Accuracy : 0.6836
##                  95% CI : (0.6698, 0.6972)
##     No Information Rate : 0.5058
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.367
##
##  Mcnemar's Test P-Value : 0.2662
##
##             Sensitivity : 0.6703
##             Specificity : 0.6966
##          Pos Pred Value : 0.6834
##          Neg Pred Value : 0.6837
##              Prevalence : 0.4942
##          Detection Rate : 0.3313
##    Detection Prevalence : 0.4847
##       Balanced Accuracy : 0.6834
##
##        'Positive' Class : no
##
```

Finally, the most relevant variables for the Random Forest process are: if the youth had children, if they worked while studying and if they were married.

```
## ranger variable importance
##
##                  Overall
## work_while.study 100.000
## child             94.349
## age               65.458
## married           29.548
## financial_sit      4.006
## sex                0.000
```

Conclusion

In the project proposal I establish as "success" if I can provide evidence of which is the youth profile that successfully enroll the job market and which is the weight of each characteristic

in the probability of success. For example, if years of education influence the probability to find a job, in which proportion, in average the impact is.

What I can conclude is that I partially succeed. From the failure side, I couldn't use variables that I considered relevant at the beginning such as years of education or if the youth live in urban or rural ares, however, this is because of the quality of the data. As well, I expected to be able to provide the weight of influence of each independent variable in the probability of get a job but I failed.

However, from the success side, the model I proposed, after Random Forest process, can predict successfully the 67% of correct classification as employed or unemployed. In this sense, even though I couldn't provide a weight, I am able to say that the most predictive variables are if the youth have children, if they worked while studying, their age and if they are married.

In terms of public policy interventions, which was my main interest, I don't get very relevant conclusions, but interesting confirmations from previous studies can be provided, nonetheless. For example, that in the context of Sub-Saharan countries, having a job while studying increase the chance to get a job after school. This outcome can be used to support part time jobs programs for young people while studying. **Note:** Each time I knit the document, all

the ML results change and the best model change as well as the ROC, confusion matrix, etcetera. Just to let you know that my results and conclusions are over a knit version different to this.

References:

- World Bank. 2018. Integrated Youth Employment Programs – A Stock-take of Evidence on what work in Youth Employment programs, p. 24
- Goldin, N. & M. Hobson with P. Glick, M. Lundberg, S. Puerto (S4YE). 2015. "Toward Solutions for Youth Employment: A Baseline for 2015." Solutions for Youth Employment, Washington D.C
- OECD (2018), The Future of Rural Youth in Developing Countries: Tapping the Potential of Local Value Chains, OECD Publishing, Paris.