HACKATON PRESENTATION

Team Sustainability

Marine Jouvin

Jean-Philippe Kouadio

Oumaïma Boukamel

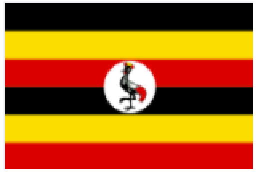
Who are we ?

Jean-Philippe Kouadio: Data Scientist, based in Abidjan, Côte d'Ivoire

Marine Jouvin: PhD in Development Economics, based in Bordeaux, France

Oumaïma Boukamel: M&E Manager, based in Bordeaux, France

Our Scope



Analysis focusing on Uganda households.

Analysis based on a sampe of 2225 households surveyed by the *World Bank* and the *Ugandan Office of Statistics.*

Uganda is located in East Africa and has known pretty severe lockdown measures during COVID-19.

Area	
• Total	241,038 km^2 (93,065 sq mi) (79th)
• Water (%)	15.39
Population	
• 2018 estimate	▲ 42,729,036[5][6] (35th)
• 2014 census	▲ 34,634,650[7]
• Density	157.1/km^2 (406.9/sq mi)
GDP (PPP)	2019 estimate
• Total	$102.659 billion[8]
• Per capita	$2,566[8]
GDP (nominal)	2019 estimate
• Total	▲ $30.765 billion[8]
• Per capita	▲ $956[8]

Source: Wikipédia

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

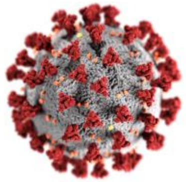
What is vulnerability ?

"Vulnerability is the inability to resist a hazard or to respond when a disaster has occurred. For instance, people who live on plains are more vulnerable to floods than people who live higher up."

unisdr.org

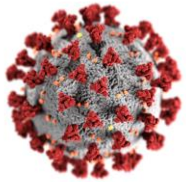
Our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Our objective

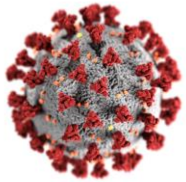
Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda

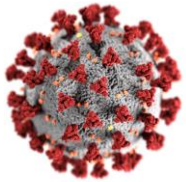


Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

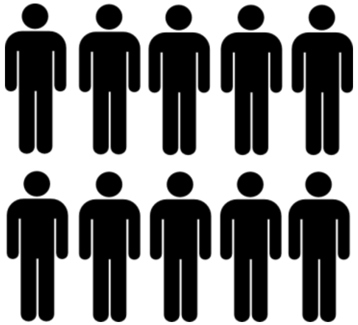
The data

World Bank Microdata Library: contains 3626 studies

The data

World Bank Microdata Library: contains 3626 studies

What we selected:



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

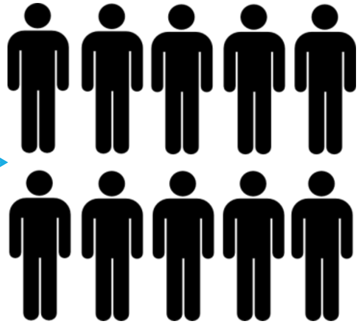
The data

World Bank Microdata Library: contains 3626 studies

What we selected:

THE WORLD BANK
IBRD • IDA

LSMS Survey 19-20 containing data on the socio economic characteristics of COVID



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

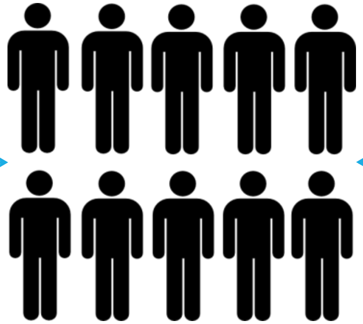
The data

World Bank Microdata Library: contains 3626 studies

What we selected:

LSMS Survey 19-20 containing data on the socio economic characteristics of COVID

High Frequency Phone survey on COVID 2020-2021 containing data on the impact and coping of COVID on households



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

The data

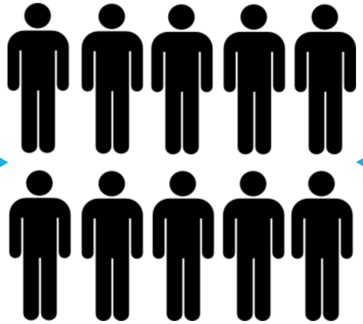
Combining both datasets enabled us to have a set of variables that we could use as « predictors » (LSMS variables) and a set of variables that we could use as « predictions » (COVID data).

World Bank Microdata Library: contains 3626 studies

What we selected:

LSMS Survey 19-20 containing data on the socio economic characteristics of COVID

High Frequency Phone survey on COVID 2020-2021 containing data on the impact and coping of COVID on households



The same sample of 2225 households in Uganda was covered by several surveys conducted by the World Bank and the Uganda Bureau of statistics

Data description

- The LSMS contains two datasets:
  - One dataset at the household level
  - One dataset at the household member level

Data description

- The LSMS contains two datasets:
  - One dataset at the household level
  - One dataset at the household member level
- The high frequency phone survey on COVID contains overall 16 datasets, but we used 8 of them:
  - The cover containing identification information
  - The household roster containing information on the household members
  - A dataset on the level of knowledge of respondents on COVID-19
  - A dataset on the behavior adopted by the respondent to cope with the pandemic
  - A dataset showing the level of access to COVID protection
  - A dataset on the impact of COVID on the crops
  - A dataset on the impact of COVID on income (it is an income level dataset meaning that there is one observation per income source)
  - A dataset on the impact of COVID on food security

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Merging the High Frequency Phone COVID Survey datasets:

- All datasets contained a unique household ID (HHID) that was used to merge all datasets

Data description

Merging the LSMS datasets:

- Both datasets contained a unique household ID (baselinehhid) that was used to merge both datasets

Merging the High Frequency Phone COVID Survey datasets:

- All datasets contained a unique household ID (HHID) that was used to merge all datasets

Merging the High Frequency Phone COVID Survey datasets:

- The dataset containing identification information on the survey also contained the LSMS household ID (baselinehhid) that unabled us to link the datasets.

Data processing and cleaning

STEP 1: Cleaning the two surveys separately

- Check duplicates
- Fix structural errors
- Outliers identification
- Rename columns to make the variables names more transparent and to avoir duplicated of variable names among the different datasets
- Validation and cross-checking

Data processing and cleaning

STEP 2: Synthetizing rosters to get one comprehensive datasets with 1 observation per household

- LSMS: Synthesis of the household member roster (total household size, indicators on education level, education level of the household head, proportion of litterate household members, number of household member per age range and gender etc…)

Find | Next | Prev | All | Replace | Replace | All

In selection | Match case | Whole word | Regex | Wrap

```
aggregated_roster<-clean_hh_roster_lsms%>%
  group_by(hhid)%>%
  summarise(hh_size=sum(count),
            f0_14=mean(f0_14),
            f15_64=mean(f15_64),
            f65=mean(f65),
            m0_14=mean(m0_14),
            m15_64=mean(m15_64),
            m65=mean(m65),
            married=mean(married,na.rm=TRUE),
            form_married=mean(form_married,na.rm=TRUE),
            nev_married=mean(nev_married,na.rm=TRUE),
            literacy=mean(literacy,na.rm=TRUE),
            work=mean(work,na.rm=TRUE),
            primary_age=sum(primary*age_5_11),
            secondary_age=sum(secondary*age_12_17),
            tertiary_age=sum(tertiary*age_18),
            primary=sum(primary),
            secondary=sum(secondary),
            tertiary=sum(tertiary),
            tot_5_11=sum(age_5_11),
            tot_12_17=sum(age_12_17),
            tot_18=sum(age_18))


colnames(aggregated_roster)[1]<-"baseline_hhid"
aggregated_roster$prop_primary<-aggregated_roster$primary/aggregated_roster$hh_size
aggregated_roster$prop_secondary<-aggregated_roster$secondary/aggregated_roster$hh_size
aggregated_roster$prop_tertiary<-aggregated_roster$tertiary/aggregated_roster$hh_size
aggregated_roster$prop_educated<-aggregated_roster$prop_primary+aggregated_roster$prop_secondary+aggregated_roster$prop_tertiary
summary(aggregated_roster$prop_primary)
summary(aggregated_roster$prop_secondary)
summary(aggregated_roster$prop_tertiary)
```

Data processing and cleaning

STEP 2: Synthetizing rosters to get one comprehensive datasets with 1 observation per household

- COVID Survey: The roster dataset contained variables with one line per household*type of income source. We synthetized the dataset in order to get for each household total the number of income sources, the proportion of income sources completely lost due to COVID and the proportion of income sources reduced due to COVID.

```
#Income data aggregation per household

income_summary<-income_loss_covid_r1[income_loss_covid_r1$income_source_lastmonths==1,]
income_summary$counting<-rep(1,nrow(income_summary))
income_summary$reduced<-rep(0,nrow(income_summary))
income_summary$no_income<-rep(0,nrow(income_summary))
income_summary$reduced[income_summary$income_evolution==3]<-1
income_summary$no_income[income_summary$income_evolution==4]<-1

income_summary<-income_summary%>%
  group_by(HHID)%>%
  summarise(nb_income=sum(counting),nb_reduced=sum(reduced),nb_noincome=sum(no_income))


income_summary$fq_reduced<-income_summary$nb_reduced/income_summary$nb_income
income_summary$fq_noincome<-income_summary$nb_noincome/income_summary$nb_income
income_summary$total_loss<-rep(NA,nrow(income_summary))
income_summary$reduction<-rep(NA,nrow(income_summary))
```

Multiple correspondence analysis (MCA)

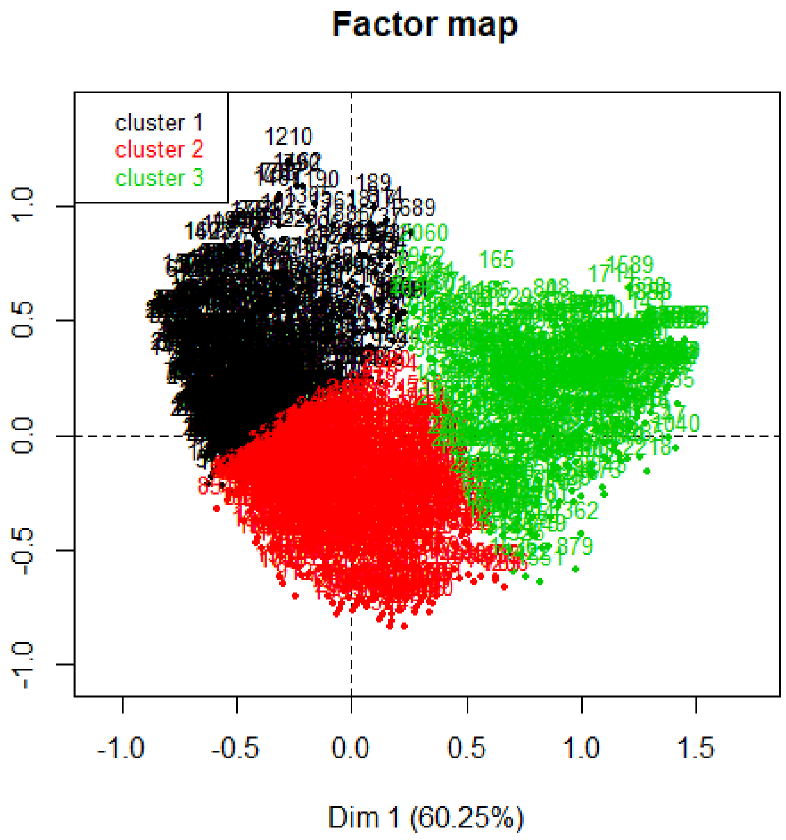
- **Objective** : to segregate households by level of vulnerability

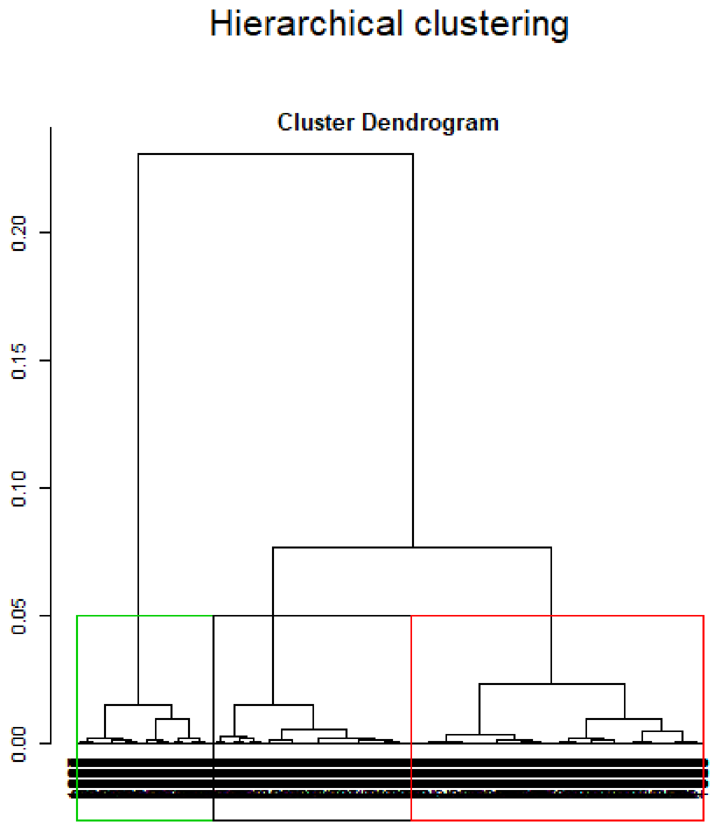
- **Method** : We rely on a MCA analysis (as we used only categorical variables), followed by a hierarchical ascending classification (HAC) consolidated by the k-means method.

- **Variables used for segmentation** :
  - **Housing** : Materials of the walls, floor and roof of the house, access to electricity, water and toilets.
  - **Assets** : Possession of a cellphone, a refrigerator, a motorcycle.
  - **Farming information** : possession of land and crop, and livestock ownership.
  - **Income** : income of the household.
  - **Household composition** : number of persons in the household, education of the household head.

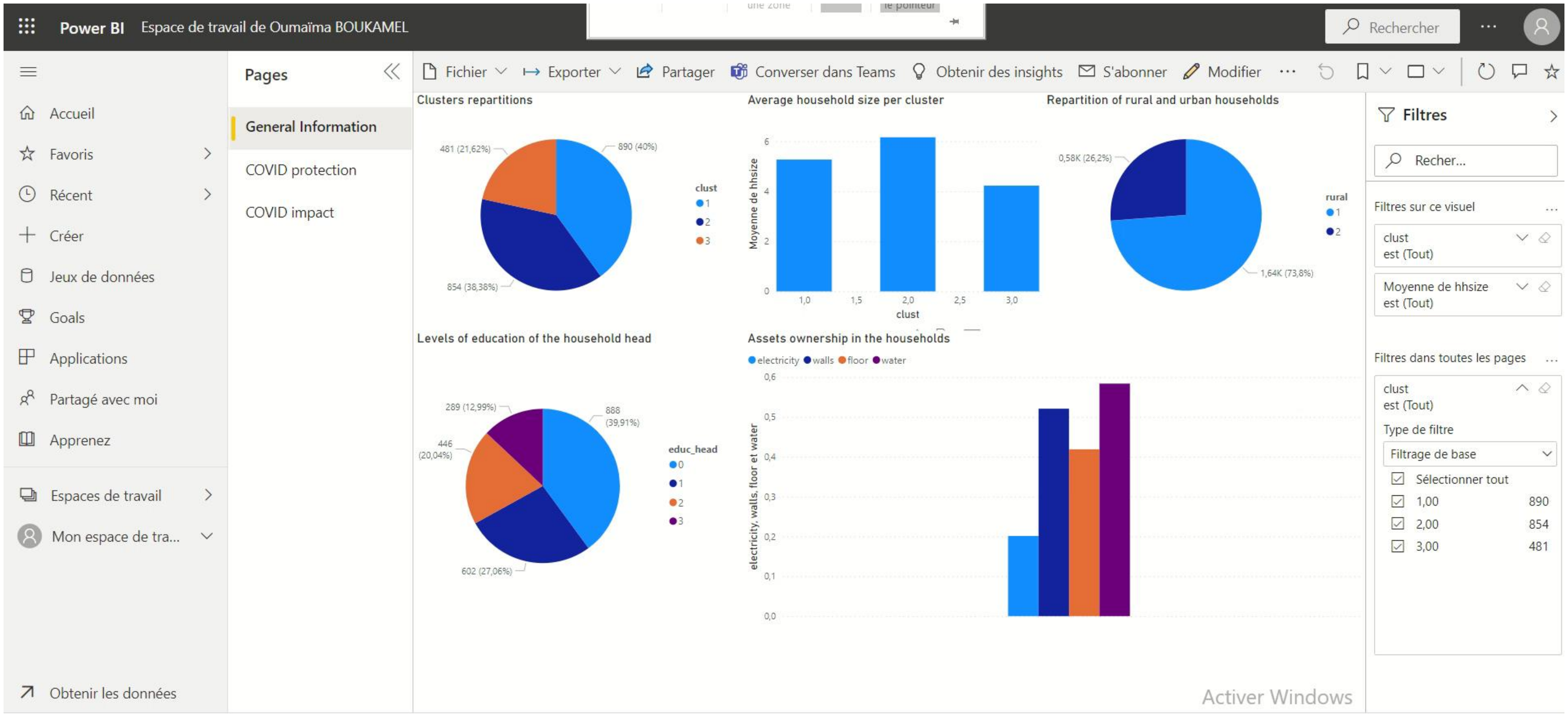
Multiple correspondence analysis (MCA)

- **Findings :** The MCA and the ACH result in the classification of households into 3 distinct groups, which explains 68% of the inter-household variance.
  - Class 1 : Poor rural households
  - Class 2 : Vulnerable rural households
  - Class 3 : Urban, less vulnerable, households





Data visualization per cluster



Data visualization per cluster

STEP 1: Import of the the data cleaning and some processing in power BI through an R script

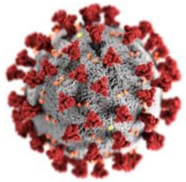
STEP 3: Adding the variable *clust* as a filter so that the user can filter the data per cluster

STEP 2: Building the visualisations on 3 thematics:

- General characteristics of the households
- COVID-19 protection characteristics
- Impact of COVID-19 on the household

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Methodology: setting-up classification models

- **Naive Bayes (with Rstudio)**

STEP 1: Import and load packages

Import and load the following packages e1071, caTools, caret

STEP 2: Split the dataset in 2 datasets (split ratio = 0.7), using sample.split. One dataset will be the **training** dataset, the other one will be the **test** dataset.

```
split<-sample.split(c(1:nrow(M)),SplitRatio=0.7)
train_cl<-subset(M,split==TRUE)
test_cl<-subset(M,split==FALSE)
```

STEP 3: Scaling of the datasets to « smooth » the data using the function scale

Methodology: setting-up classification models

- **Naive Bayes (with Rstudio)**

STEP 4: Setting seeds (set.seed(120))

STEP 5: Applying the naiveBayes fonction and generating the classifier using the training dataset

```
classifier_cl <- naiveBayes(fs_vulnerability ~ ., data=train_cl)
classifier_cl
```

STEP 6: Predicting on the test data

```
# Predicting on test data
y_pred <- predict(classifier_cl,newdata=test_cl)
```

STEP 7: Model evaluation (using the confusion matrix to compare the predictions with the actual values)

Methodology: setting-up classification models

- **Decision trees (with Rstudio)**

STEP 1: Import and load packages (DAAG, party, rpart, rpart.plot,mlbench, caret, pROC, tree)

STEP 2: Converting the « prediction category » in factors (with as.factor) and setting seeds (set.seed(1234))

STEP 3: Split the dataset in 2 datasets (split ratio = 0.5). One dataset will be the **training** dataset, the other one will be the **test** dataset.

```
ind<-sample(2,nrow(M),replace=T, prob = c(0.5,0.5))
train<- subset(M,ind==1)
test<-subset(M,ind==2)
```

Methodology: setting-up classification models

- **Decision trees (with Rstudio)**

STEP4: Tree classification

```
# Tree classification

tree <-rpart(fs_vulnerability ~., data=train)
rpart.plot(tree,box.palette="blue")

printcp(tree)

rpart(formula = fs_vulnerability ~., data=train)

plotcp(tree)
```

STEP 5: Testing the prediction model on the test data and comparing the outputs to the actual categories

STEP 6: Model evaluation with the confusion matrix (confusionMatrix function)

Methodology: setting-up classification models

- **K-NN (with Rstudio)**

STEP 1: Inputing relevant values to NA as the K-NN model does not work if the data contains empty values

STEP 2: defining a normalization function and run the normalization on the predictor

```
## the normalization function is created

nor <-function(x){(x-min(x)/max(x)-min(x))}


## Run normalization on the predictors

M_norm <- data.frame(lapply(M[,-1],nor))
```

Methodology: setting-up classification models

- **K-NN (with Rstudio)**

STEP 3: Split the dataset in 2 datasets (split ratio = 0.8). One dataset will be the **training** dataset, the other one will be the **test** dataset.

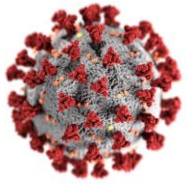
STEP 4: Run the K-NN function

```
##run knn function
pr <- knn(M_train, M_test, cl=M_target_category)
```

STEP 5: Model evaluation with the confusion matrix

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Model 1: Identifying income vulnerability

Defining the categories

Category	Proportion of income sources lost range	Number of households in this category
The household has lost all their income sources during the pandemic	=1	123
The household has lost less than 50% of their income sources during the pandemic	<0.5	117
The household has lost more than 50% of their income sources during the pandemic	>=0.5	292
The household has lost none of their income sources during the pandemic	=0	1693

The proportion of income sources completely lost was calculated from the income source roster of the High Frequency Phone Survey on COVID-19, that was cleaned and aggregated.

Model 1: Identifying income vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, primary_head, secondary_head, tertiary_head

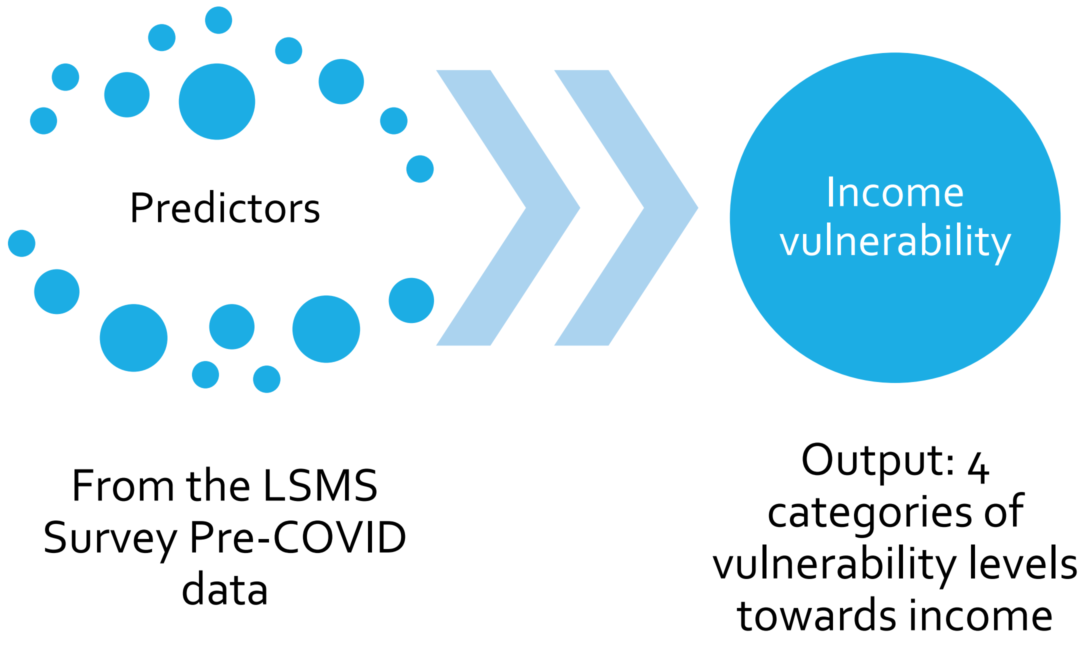
From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels towards income

Model 1: Identifying income vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**



Model 1: Identifying income vulnerability

K-NN Classification results

```
Statistics by Class:

                     Class: The household lost all their income sources
Sensitivity                                                    0.142857
Specificity                                                    0.948113
Pos Pred Value                                                 0.120000
Neg Pred Value                                                 0.957143
Prevalence                                                     0.047191
Detection Rate                                                 0.006742
Detection Prevalence                                           0.056180
Balanced Accuracy                                              0.545485
                     Class: The household lost less than 50% of their income sources
Sensitivity                                                                 0.095238
Specificity                                                                 0.941038
Pos Pred Value                                                              0.074074
Neg Pred Value                                                              0.954545
Prevalence                                                                  0.047191
Detection Rate                                                              0.004494
Detection Prevalence                                                        0.060674
Balanced Accuracy                                                           0.518138
                     Class: The household lost more than 50% of their income sources
Sensitivity                                                                  0.13462
Specificity                                                                  0.89313
Pos Pred Value                                                               0.14286
Neg Pred Value                                                               0.88636
Prevalence                                                                   0.11685
Detection Rate                                                               0.01573
Detection Prevalence                                                         0.11011
Balanced Accuracy                                                            0.51387
                     Class: The household lost no income sources
Sensitivity                                                0.7892
Specificity                                                0.2872
Pos Pred Value                                             0.8052
Neg Pred Value                                             0.2673
Prevalence                                                 0.7888
Detection Rate                                             0.6225
Detection Prevalence                                       0.7730
Balanced Accuracy                                          0.5382
```

Naive Bayes classification results

```
Statistics by Class:

                     Class: The household lost all their income sources
Sensitivity                                                     0.07500
Specificity                                                     0.97872
Pos Pred Value                                                  0.90000
Neg Pred Value                                                  0.29299
Prevalence                                                      0.71856
Detection Rate                                                  0.05389
Detection Prevalence                                            0.05988
Balanced Accuracy                                               0.52686
                     Class: The household lost less than 50% of their income sources
Sensitivity                                                                 0.071429
Specificity                                                                 0.953674
Pos Pred Value                                                              0.093750
Neg Pred Value                                                              0.938679
Prevalence                                                                  0.062874
Detection Rate                                                              0.004491
Detection Prevalence                                                        0.047904
Balanced Accuracy                                                           0.512551
                     Class: The household lost more than 50% of their income sources
Sensitivity                                                                 0.250000
Specificity                                                                 0.893939
Pos Pred Value                                                              0.027778
Neg Pred Value                                                              0.989933
Prevalence                                                                  0.011976
Detection Rate                                                              0.002994
Detection Prevalence                                                        0.107784
Balanced Accuracy                                                           0.571970
                     Class: The household lost no income sources
Sensitivity                                                0.8333
Specificity                                                0.2283
Pos Pred Value                                             0.2195
Neg Pred Value                                             0.8403
Prevalence                                                 0.2066
Detection Rate                                             0.1722
Detection Prevalence                                       0.7844
Balanced Accuracy                                          0.5308
```

Model 1: Identifying income vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.4332, 0.5102)
K-NN	(0.6031, 0.6938)

We decided to go for the K-NN based on the accuracy confidence interval and based on the comparison of the sensitivity and specificity of the category « The household lost all their income sources » which is the category that we want to determine in priority.

Model 1: Identifying income vulnerability

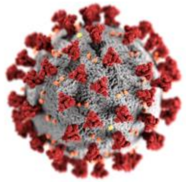
Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.4332, 0.5102)
K-NN 	(0.6031, 0.6938)

We decided to go for the K-NN based on the accuracy confidence interval and based on the comparison of the sensitivity and specificity of the category « The household lost all their income sources » which is the category that we want to determine in priority.

Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Model 2: Identifying food security vulnerability

Figure 4. Actual Example—Calculating a Household CSI Index Score

In the past 7 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	Raw Score	Severity Weight	Weighted Score = Frequency X weight
(Add each behavior to the question)			
a. Rely on less preferred and less expensive foods?	5	1	5
b. Borrow food, or rely on help from a friend or relative?	2	2	4
c. Purchase food on credit?	1	2	2
d. Gather wild food, hunt, or harvest immature crops?	0	4	0
e. Consume seed stock held for next season?	0	3	0
f. Send household members to eat elsewhere?	1	2	2
g. Send household members to beg?	0	4	0
h. Limit portion size at mealtimes?	7	1	7
i. Restrict consumption by adults in order for small children to eat?	2	2	4
j. Feed working members at the expense of non-working members?	0	2	0
k. Reduce number of meals eaten in a day?	5	2	10
l. Skip entire days without eating?	0	4	0
TOTAL HOUSEHOLD SCORE	Sum down the totals for each individual strategy		**34**

- This CSI index Score was developed under the framework of collaborative research project, implemented by WFP and CARE in Kenya, with financial support of the UK Department for International Development via WFP, The Bill and Melinda Gates Foundation, and CARE-USA.
- Among the items described on the item described on the left the High Frequency Phone Survey on COVID contains the items a,k,h and l.
- We used this Score definition to set the ponderations of an index we designed in order to assess the food insecurity levels of the households during COVID
- Based on this index we defined 4 categories of households based on their food insecurity level: "Not vulnerable", "Moderately vulnerable", "Very vulnerable", "Severely vulnerable".

Model 2: Identifying food security vulnerability

Defining the index

Question	Variable	Severity	CSI Index Score equivalent	Ponderation
Were you or any other adult in your household were worried about not having enough food to eat because of lack of money or other resources?	fs_worried	1		1/14
You, or any other adult in your household, were unable to eat healthy and nutritious/preferred foods because of a lack of money or other resources?	fs_healthy	1	a. Rely on less preferred and less expensive food	1/14
You, or any other adult in your household, ate only a few kinds of foods because of a lack of money or other resources?	fs_few	1		1/14
You, or any other adult in your household, skipped meals because of a lack of money or other resources?	fs_skip	2	k. Reduce number of meals eaten in a day	2/14
You, or any other adult in your household, ate less than you thought you should because of a lack of money or other resources?	fs_less	1	h. Limit portion size at meal time	1/14
Your household ran out of food because of a lack of money or other resources?	fs_ranout	2		2/14
You, or any other adult in your household, were hungry but did not eat because there was not enough money or other resources for food?	fs_hungry	2		2/14
You, or any other adult in your household, went without eating for a whole day because of a lack of money or other resources?	fs_day	4	l. Skipped entire days without eathing	4/14

Model 2: Identifying food security vulnerability

Defining the categories

Category	Index range	Number of households in this category
Not vulnerable	Index==0	563
Moderately vulnerable	Index in]0,0.28[	639
Very vulnerable	Index in [0.28, 0,5[	380
Severely vulnerable	Index in >=0,5	643

The categories were defined to ensure that the households who checked an item with a severity score equal to 4 or two items with a severity score equal to 2 (hence with an index superior or equal to 2/7) were in the category very vulnerable or severely vulnerable.

Model 2: Identifying food security vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, primary_head, secondary_head, tertiary_head

From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 2: Identifying food security vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**
- **Decision Trees**

From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 2: Identifying food security vulnerability

Naives Bayes

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4984	0.33333	0.6923	0.0000
Specificity	0.6073	0.87875	0.7175	1.0000
Pos Pred Value	0.3383	0.48469	0.5011	NaN
Neg Pred Value	0.7504	0.79393	0.8505	0.8327
Prevalence	0.2871	0.25492	0.2907	0.1673
Detection Rate	0.1431	0.08497	0.2013	0.0000
Detection Prevalence	0.4231	0.17531	0.4016	0.0000
Balanced Accuracy	0.5529	0.60604	0.7049	0.5000

Decision Tree

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4984	0.33333	0.6923	0.0000
Specificity	0.6073	0.87875	0.7175	1.0000
Pos Pred Value	0.3383	0.48469	0.5011	NaN
Neg Pred Value	0.7504	0.79393	0.8505	0.8327
Prevalence	0.2871	0.25492	0.2907	0.1673
Detection Rate	0.1431	0.08497	0.2013	0.0000
Detection Prevalence	0.4231	0.17531	0.4016	0.0000
Balanced Accuracy	0.5529	0.60604	0.7049	0.5000

K-NN

Statistics by Class:

	Class: Moderately vulnerable	Class: Not vulnerable	Class: Severely vulnerable	Class: Very vulnerable
Sensitivity	0.4267	0.20000	0.4789	0.34375
Specificity	0.7095	0.74719	0.8092	0.89529
Pos Pred Value	0.4267	0.16667	0.5397	0.35484
Neg Pred Value	0.7095	0.78698	0.7687	0.89062
Prevalence	0.3363	0.20179	0.3184	0.14350
Detection Rate	0.1435	0.04036	0.1525	0.04933
Detection Prevalence	0.3363	0.24215	0.2825	0.13901
Balanced Accuracy	0.5681	0.47360	0.6440	0.61952

Model 2: Identifying food security vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.3345, 0.4091)
K-NN	(0.2536, 0.3792)
Decision trees	(0.4001, 0.459)

Based on the Accuracy CI we decided to go with the Decision tree model.

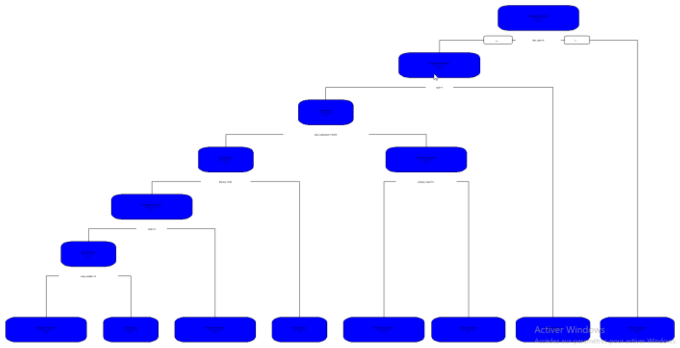
Model 2: Identifying food security vulnerability

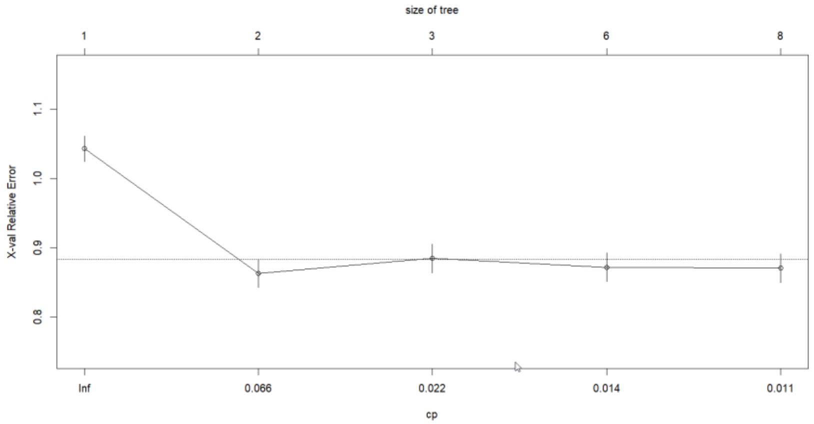
Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.3345, 0.4091)
K-NN	(0.2536, 0.3792)
Decision trees 	(0.4001, 0.459)

Based on the Accuracy CI we decided to go with the Decision tree model.

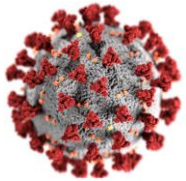
Decision tree visuals





Back to our objective

Understanding household's vulnerability to COVID's consequences in Uganda



Identifying the most vulnerable households towards loss of income due to the COVID pandemic:
What are the household profiles that are the most likely to lose one or several of their income sources due to COVID?

Identifying the most vulnerable households towards food security:
What are the household profiles that are most likely to face food insecurity due to COVID ?

Identifying the most vulnerable households towards education:
What are the household profiles in which children are more likely to drop school due to the pandemic ?

Model 3: Identifying education access vulnerability

Defining the categories

Category	Value of the variable children_school_covid	Number of households in this category
The children of the households have continued learning activities after the pandemic	=1	1034
The children of the households have stopped learning activities after the pandemic	=2	699

Model 3: Identifying education access vulnerability

Within the LSMS dataset we chose the following predictors:

- Rural, roof, floor, walls, toilet,water,rooms,elect,tv,radio,refrigerator,land_tot,land_cultivated, rent, remit, assist, crop, crop_number, cash_crop, sell_crop, fies_mod, fies_sev, hh_size, adulteq, literacy, work, prop_primary, prop_secondary, prop_tertiary

From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels towards education

Model 3: Identifying education access vulnerability

We tested 3 classification methodologies in order to select the most performant one:

- **Naives Bayes Classifier**
- **K-NN**

From the LSMS Survey Pre-COVID data

Output: 4 categories of vulnerability levels to food insecurity

Model 3: Identifying education access vulnerability

Naive Bayes

```
    M_test_category
pr     1   2
  1 114  75
  2  92  66

               Accuracy : 0.5187
                 95% CI : (0.4648, 0.5724)
    No Information Rate : 0.5937
    P-Value [Acc > NIR] : 0.9980

                  Kappa : 0.0211

 Mcnemar's Test P-Value : 0.2157

            Sensitivity : 0.5534
            Specificity : 0.4681
         Pos Pred Value : 0.6032
         Neg Pred Value : 0.4177
             Prevalence : 0.5937
         Detection Rate : 0.3285
   Detection Prevalence : 0.5447
      Balanced Accuracy : 0.5107

       'Positive' Class : 1
```

K-NN

```
Confusion Matrix and Statistics

   y_pred
      1   2
  1 156 160
  2  68 136

               Accuracy : 0.5615
                 95% CI : (0.5177, 0.6047)
    No Information Rate : 0.5692
    P-Value [Acc > NIR] : 0.6555

                  Kappa : 0.1485

 Mcnemar's Test P-Value : 1.674e-09

            Sensitivity : 0.6964
            Specificity : 0.4595
         Pos Pred Value : 0.4937
         Neg Pred Value : 0.6667
             Prevalence : 0.4308
         Detection Rate : 0.3000
   Detection Prevalence : 0.6077
      Balanced Accuracy : 0.5779

       'Positive' Class : 1
```

Model 3: Identifying education access vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.5177, 0.6047)
K-NN	(0.4878, 0.5951)

Naive Bayes has a better accuracy CI but K-NN seems to detect better the cases of households whose children has stopped learning during COVID. In the logic of detecting vulnerability this is our priority: we will thus choose the K-NN model.

Model 3: Identifying education access vulnerability

Testing different classification methodology

Classification methodology	Accuracy CI
Naïve-Bayes	(0.5177, 0.6047)
K-NN 	(0.4878, 0.5951)

Naive Bayes has a better accuracy CI but K-NN seems to detect better the cases of households whose children has stopped learning during COVID. In the logic of detecting vulnerability this is our priority: we will thus choose the K-NN model.

Integrated solution

- Combination of 3 models in order to predict the different categories regarding income, food security and education in which a given household is likely to fall in.
- <u>**Conclusion:**</u>
  - **For income and education access: K-NN model will be used**
  - **For food security: Decision tree model will be used**

Next step*: write an integrated script that takes any socio-economic dataset containing the predictors as arguments and that returns the categories predicted for the household income, education access and food security evolution with COVID-19.*

Application : Context

- TOUTON SA is a company specialized in soft commodities. The sustainability department of TOUTON manages several sustainability projets in sourcing countries (including Uganda, Ghana, Côte d'Ivoire, Kenya, Nigeria and Madagascar) aiming at helping farmers improving their income and livelihoods and requiring large scale data collection.
- TOUTON has collected data on a sample of 304 coffee farmers in Uganda on their livelihoods and agricultural practices. Several variables included in this survey have been used as predictors for our different prediction models.
- Therefore, with the consent of TOUTON SA, we have applied our different models that we developped with open source data to their coffee farmers datasets in order to assess their vulnerability to COVID regarding food security and their access to education.

Application : Cleaning and processing

STEP0: Getting all parties consent to use the data for visualisation only

STEP 1: Retrieving the predictors from the coffee farmer survey in Uganda

STEP 2: Cleaning the data and replacing missing values (using extrapolations)

STEP 3: Import the dataset in the integrated script and applying the 2 predicting models on income, food security and education access to the dataset

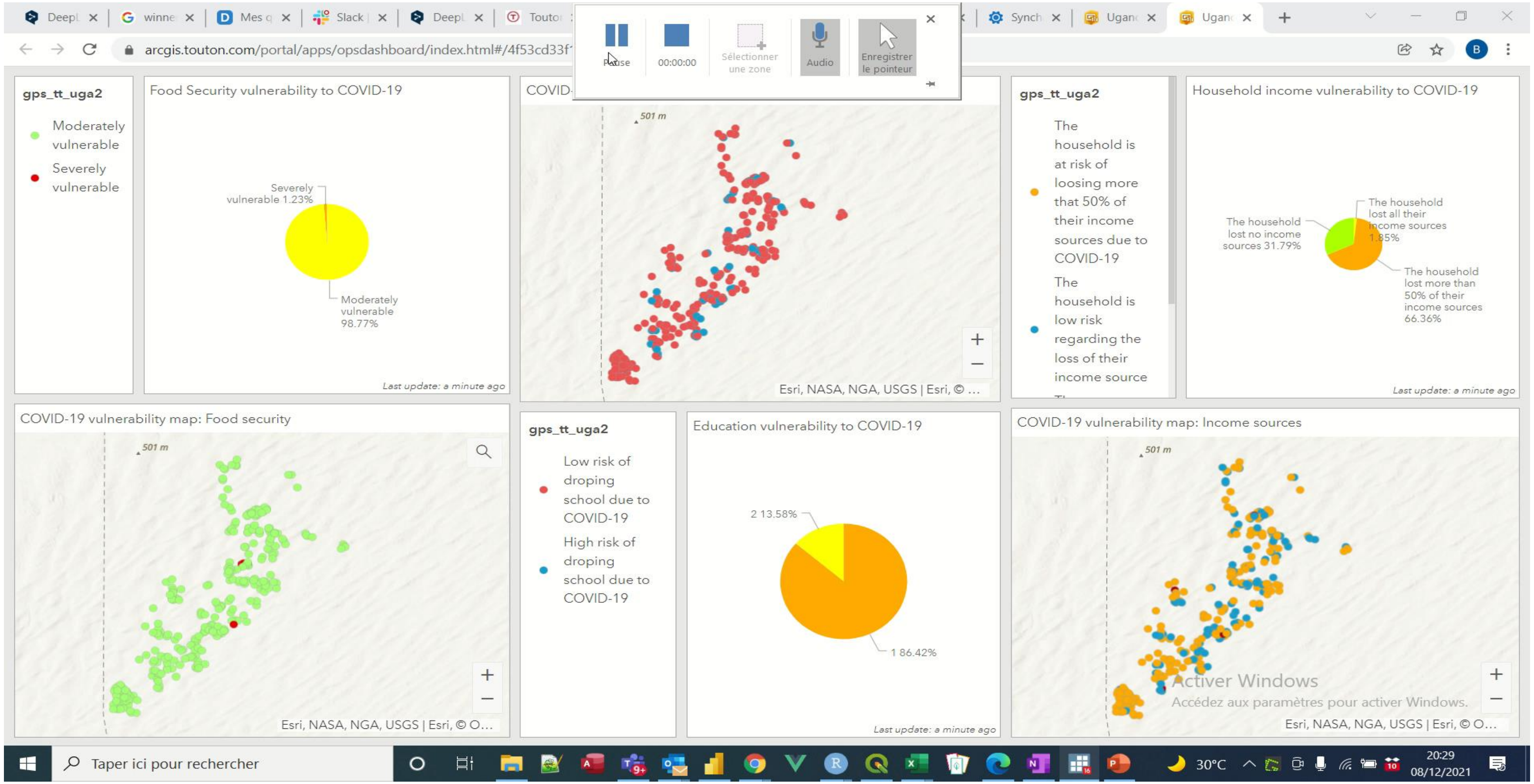
STEP 4: Creating a dataset containing the farmer ID as well as the 3 predictions. This dataset is the prediction dataset.

STEP 5: Merging the geospatial data on farmers with the « prediction dataset ».

STEP 6: Importing the data in Arcgis enterprise

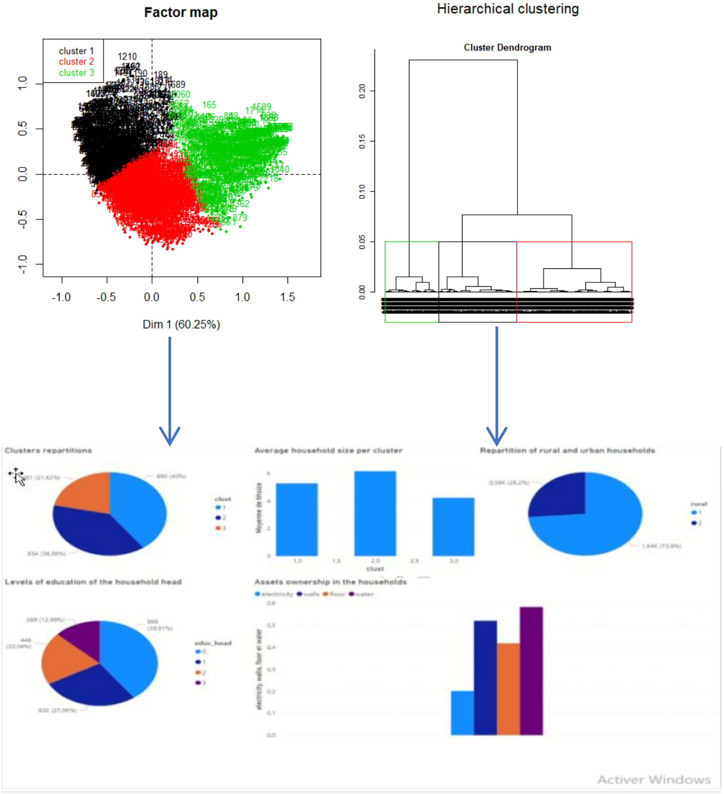
STEP 7: Building a « Vulnerability map dashboard » to visualise the results

Application : Visualizing coffee farmers that are the most vulnerable to COVID consequences

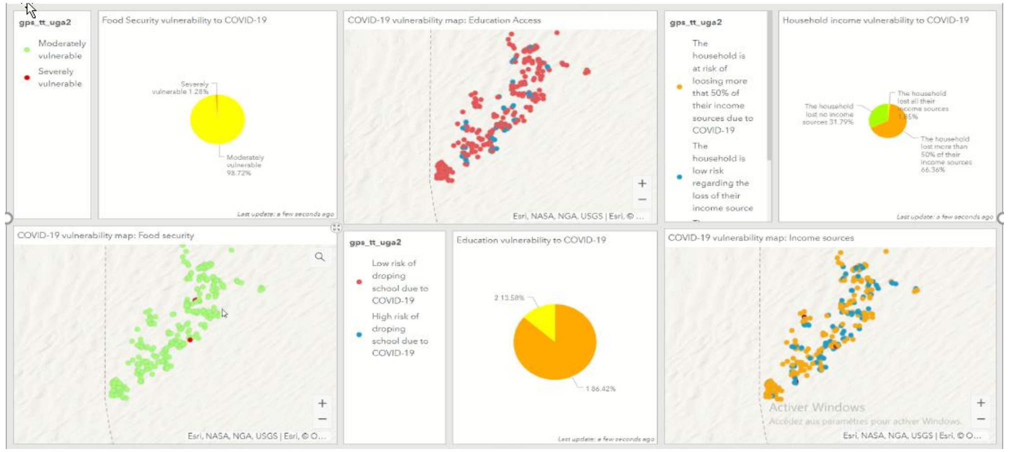


Conclusion: Our solution

A statistical segmentation to better understand the impact of a household socio-economic characteristics on their vulnerability to COVID-19 and their consequences.



A integrated prediction model in order to assess the vulnerability of households to COVID-19 regarding their income, food security and education access



Conclusion

What we can improve:

i) The World Bank's microdata catalogue contains similar datasets collected from households in Malawi, Ethiopia, Nigeria, Cambodge etc. The analysis could therefore be run on a larger set of data and thus be more accurate

ii) Improving the segmentation dashboard with more data, variables and correlation studies

iii) Test with different predictors to see if get better accuracies

iv) Further classification models (such as logistic regression or random forest – especially for the ones in which the decision tree worked well) could be tested

v) As other surveys are available, it would be possible to get other kinds on data on the households to run the analytics

vi) Automate the analysis by developping a function that automatically tests several models and choses the best model based on performance criteria to define

vii) The survey observed evolution of the socio-economic characteristics of the households based on the household's declaration: therefore this is not an observed evolution based on data from one year to another. Based on other datasets collected by the world bank in the future we could proceed this way for further analysis

Annex 1: Deliverables description

Script	What's in there?
Data_cleaning	Data cleaning and processing
Data_exploration	First exploration of the data
Classification_education_testing	Testing of classifications on education
Classification_food_security_testing	Testing of classifications on food security
Classification_income_testing	Testing of classifications on income
Dashboard_Segmentation_Script	Script to import the data in power BI for the segmentation dashboard
Integrated_Prediction_Script	Integrated script combining all prediction model selected and applied on the TOUTON data
Segmentation_FAMD_HCPC	Script segmentation

Annex 2: Other files

- Variable_Dictionary contains the variable signification
- Dashboard_Hackaton is the power BI dashboard build based on the segmentation
- All the data used can be found in the folder Data

Annex 3: Data references

Data used to train the algorithm:

- LSMS dataset: https://microdata.worldbank.org/index.php/catalog/4183
- High Frequency Phone Survey on COVID-19: https://microdata.worldbank.org/index.php/catalog/3765

Data on which the model was applied:

- Uganda Socio-Economic Survey Coffee farmers: Touton Property