🎏

**Group assignment**

# What’s expected?

**Business understanding** Make sure you understand the use case and the different features (refer to **STATS406 ‐ Assignment metadata info.pdf**)

**Data merging** Merge the 3 **datasets** into one dataframe

*Python (using pandasql)*

**Data cleaning** Handle missing values, encoding errors, outliers

*Python*

**Data transformation** Ensure your data is in a suitable format for the classification models

& select the relevant features for modelling

*Python and/or Rapid Miner*

**Data exploration** Create visualisation and look for patterns

*Rapid Miner*

**Data modelling** Train classifcation models and select the best performing model

*Rapid Miner*

**Analyse your model** Look at features importance to create some recommandations

*Rapid Miner*

#### Create predictions on the

**unlabelled data** Make the required transformation and use your model to create the

predictions file

*Python and/or Rapid Miner*

# Deliverables

* 1. **report** (in PDF, max. 8 pages) containing at least:
     1. A **data preparation part** where you explain how you prepared and cleaned the data. This should include the following points
        1. **Data merging**: explain how you merge the three datasets using python (we recommend using pandasql in python for this section).
        2. **Data cleaning**: explain how you handled missing values, possible encoding errors

and outliers, casting columns to the correct type, ... (using python).

* + - 1. **Data transformation**: explain how you transformed the data to ensure that it is

in a suitable format for use in the classification model (converting categorical variables to numerical ones, by using Rapid Miner). Explain also how you

selected the relevant features.

* + 1. A **data exploration part** where you explain which patterns, trends, and relationships you identified when performing data exploration. Support the explanation with visualizations

(we recommend using RapidMiner, but you can use the tool that you prefer for visualizations)

* + 1. **modeling part** where you explain how your model works. We expect that you at least

test 2 different models and explain them. We recommend using RapidMiner for the modeling part.

* + 1. An **assessment part** where you explain how you evaluated the model.
    2. A **recommendation part** where you advise, based on your findings, the NGO about how to prevent child labor in South Africa.
    3. An **appendix section** (size free) containing at least:
       1. The schematics (a screenshot, not the actual model) of your process in RapidMiner.
       2. A discussion of the most important design and analysis decision you had to make.

1. The **python code used** in a python code format (.py)1. This will be checked to make sure you wrote the python code yourself in Python and not using another tool or by copying it from another group. Every code will be run through a special code plagiarism tool, to check you did not copy

code from the other groups. **Plagiarism will be highly sanctioned**.

1. The **Rapidminer process** (.rmp) used.
2. The **cleaned data frame in CSV format** that you used as input for the RapidMiner process.
3. A **prediction file** in CSV format, where the first column will contain the IDs of unlabelled children, and the second column will contain a 0 if your model predicts that the child is not involved in child labor and a 1 if your model predicts that the child is involved in child labor. Refer to **unlabelled\_data.csv**.
4. A presentation, in **PDF**, of maximum 8 slides (without counting possible backup slides, of which the number is free but will not be considered for the grading). This presentation should show the “highlights” of your report, emphasizing business‐relevant conclusions and findings. This is the

# Data set - questions

household‐based aged 7‐17 years

data collection that took place in (Q3:2022).

Unique identifier allocated to each respondent.**(ID)**

Unique household identifier allocated to each household. **(UQNO)**

## Data set 1

One line/record is one child.

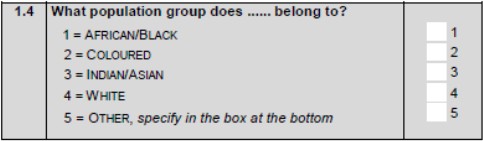
### Section 1

#### Question 1.2 ‐ Gender (Q12Gender)

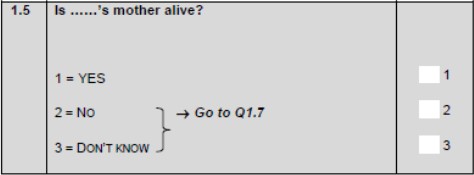
1 = Male 2 = Female

**Question 1.3 ‐ Age (Q13Age)**The instruction was to write the age in completed years to the nearest whole numbers and not in words. Thus, if a person was two years and six months, the instruction was to write the two completed years. For children aged younger than a year, the instruction was to write 000.

#### Question 1.4 ‐ Population group (Q14Population)



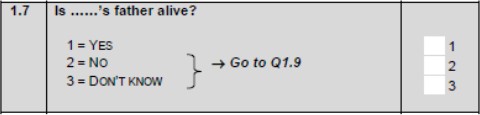
**Question 1.5 ‐ Mother alive (Q15Mothaliv)** The mother does not necessarily have to be a member of the household. That means one’s mother may be alive but living somewhere outside the selected dwelling unit. For codes 2 and 3, the instruction is to go to Question 1.7.



#### Question 1.6 ‐ Mother member of the household (Q16Mothparthh)

1 = YES 2 = NO

**Question 1.7 ‐ Father alive (Q17Fathaliv)** The father does not necessarily have to be a member of the household. That means one’s father may be alive but living somewhere outside the selected dwelling unit. For codes 2 and 3, the instruction is to go to Question 1.9.

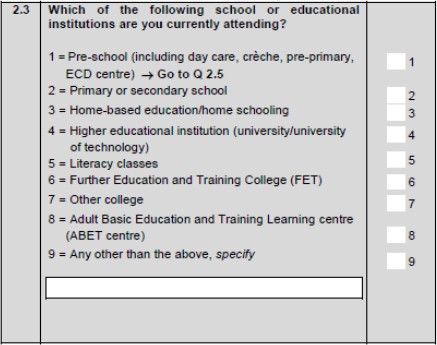


**Question 1.8 ‐ Father member of the household (Q18Fathparthh) Question 1.9 ‐ Grandparents member of the household (Q19G\_Parparthh)**

### Section 2

**Question 2.2 ‐ Currently attending school or educational institution (Q22Attend).** If the individual is not attending school or an educational institution, there is a skip to Question 2.8.

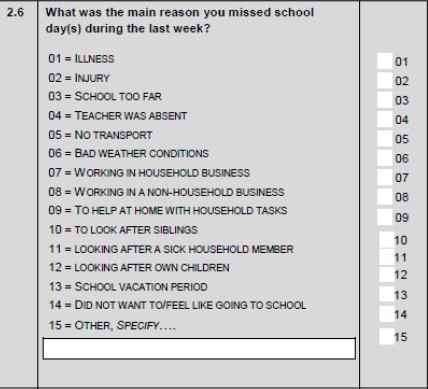
#### Question 2.3 ‐ School or educational institution (Q23Eduinst)



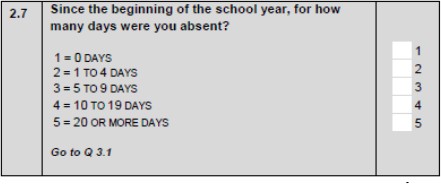
**Question 2.4 ‐ Age started Grade 1/sub‐A (Q24AGE\_GR1)**

**Question 2.5 ‐ Absent from school (Q25Miss\_Day)**It measures absenteeism from school during the school calendar week (Monday to Friday) in the week preceding the survey. Only those who respond ‘Yes’ continue to the next question. If the answer is ‘No’, skip to Question 2.7.

#### Question 2.6 ‐ Main reason absent (Q26Miss\_Rsn)

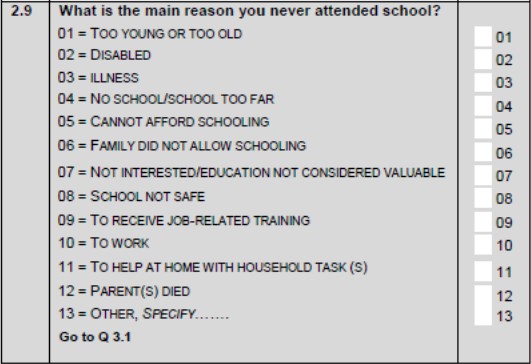


**Question 2.7 ‐ Number of days absent (Q27Daysabsent)**This question tries to find out for how many days was the child absent from school since the beginning of the school year. Note that absence excludes holidays and public holidays.



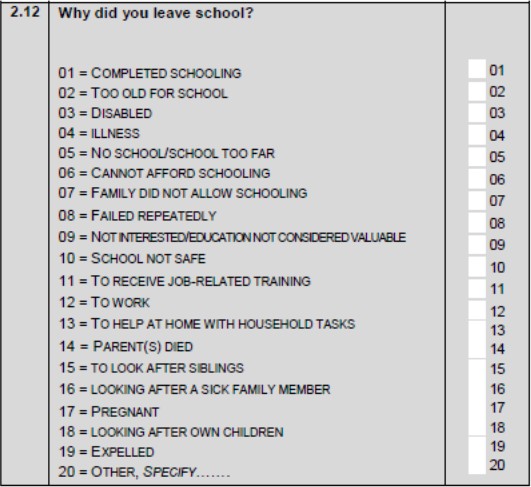
**Question 2.8 ‐ Ever attended school (Q28Everattend)**If the answer is ‘Yes’, skip to Question 2.10

**Question 2.9 ‐ Main reason never attended school (Q29Whynot)** There is a skip (Go to Q3.1) after answering this question to go to Question 3.1.



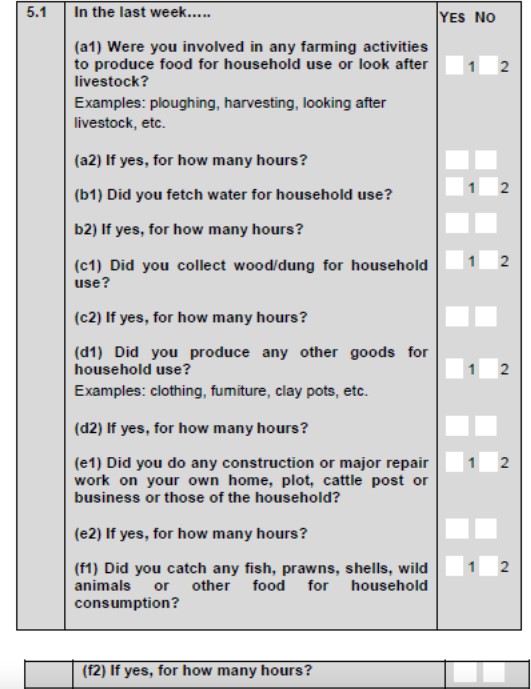
**Question 2.10 ‐ Age started Grade 1/Sub A (Q210Age\_Gr1)** The question targets the age at Grade 1, not the age the child started in.

**Question 2.11 ‐ Age left school (Q211Ageleft) Question 2.12 ‐ Reason left school (Q212Whyleft)**



### Section 5

**Question 5.1** The instruction to enumerators was that they should consider those activities that lasted for at least an hour within the seven days. In order to be certain that the categories had been answered, there should either be a 'Yes' or 'No' answer to all of them. If 'Yes' to any of these questions, the number of hours spent on the activity had to be indicated as well.



##### Do farm work (Q51AFARMWRK\_W) ‐ a1) Were you involved in any farming activities to produce food for household use or look after livestock?

***Time (Q51ATIME) (@79 2.) ‐ a2) If yes, for how many hours?***

##### Fetch water (Q51BFETCHWATER\_W) ‐ b1) Did you fetch water for household use? Time (Q51BTIME) ‐ b2) If yes, for how many hours?

***Fetch wood/dung (Q51CFETCHWOOD\_W) ‐ c1) Did you fetch wood/dung for household use?***

##### Time (Q51CTIME) ‐ c2) If yes, for how many hours?

***Produce goods (Q51DPRODHHGDS\_W) ‐ d1) Did you produce any other goods for household use?***

##### Time (Q51DTIME) ‐ d2) If yes, for how many hours?

***Do construction (Q51ECONSTRUC\_W) ‐ e1) Did you do any construction or major repair work on your own home, plot, cattle post or business or that of the household?***

##### Time (Q51ETIME) ‐ e2) If yes, for how many hours?

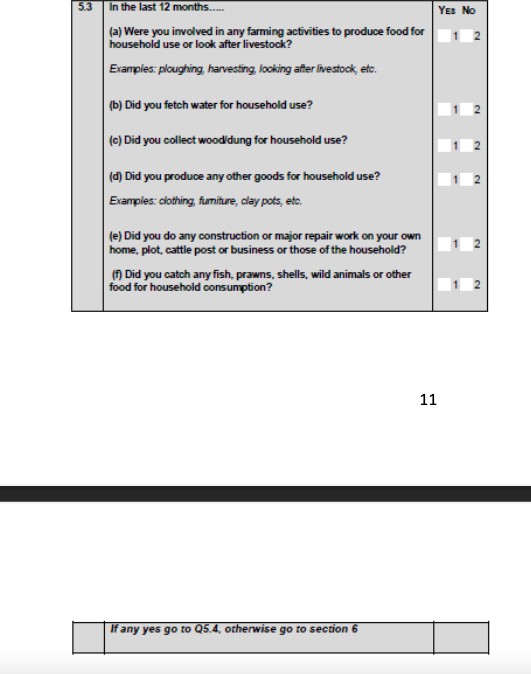
***Catch food (Q51FCATCHFOOD\_W) ‐ f1) Did you catch any fish, prawns, shells, wild animals or other food for household consumption?***

##### Time (Q51FTIME) ‐ f2) If yes, for how many hours?

**Question 5.2**to all children aged 7–17 years, and tries to find out if they have begged for money or food in public in the last week or last 12 months. In order to be certain that the categories have been answered, there should either be a 'Yes' or 'No' answer to all of them.

##### Beg last week (Q52ABEG\_LSTW) Beg last year (Q52BBEG\_LSTYR)

**Question 5.3** This question was applicable to all children aged 7–17 years, regarding their involvement in non‐ economic activities in the last 12 months. The instruction to enumerators was that they should consider those activities that lasted for at least an hour within the 12 months. In order to be certain that the categories had been answered, there should either be a 'Yes' or 'No' answer to all of them. If 'Yes' to any of these questions, the number of hours spent on the activity had to be indicated as well.



##### Do farm work (Q53AFARMWRK\_Y) ‐ a) Were you involved in any farming activities to produce food for household use or look after livestock?

***Fetch water (Q53BFETCHWATER\_Y) ‐ b) Did you fetch water for household use?***

##### Fetch wood/dung (Q53CFETCHWOOD\_Y) ‐ c) Did you fetch wood/dung for household use?

***Produce goods (Q53DPRODHHGDS\_Y) ‐ d) Did you produce any other goods for household use?***

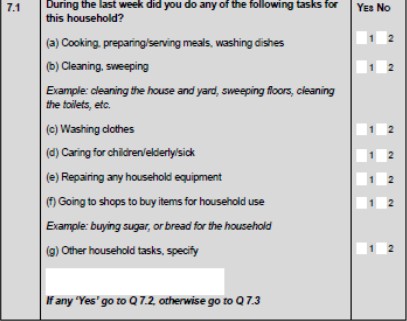
##### Do construction (Q53ECONSTRUC\_Y) ‐ e) Did you do any construction or major repair work on your own home, plot, cattle post or business or that of the household?

***Catch food (Q53FCATCHFOOD\_Y) ‐ f) Did you catch any fish, prawns, shells, wild animals or other food for household consumption?***

**Question 5.4 ‐ Hours worked (Q54TIME)** This question asks the hours the child spends on non‐market activities mentioned in Q 5.3.

### Section 7

**Question 7.1** In order to be certain that the categories had been answered, there should either be a 'Yes' or 'No' answer to all of them. If the respondent answers any 'Yes' to Question 7.1, then continue, otherwise skip to Question 7.3.



##### Cooking (Q71ACOOK) Cleaning (Q71BCLEAN)

***Washing clothes (Q71CLAUNDRY)***

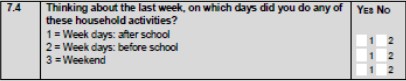
##### Caring for children (Q71DCHILDMIND) Do maintenance (Q71EMAINTENANCE)

***Went to shops (Q71FSHOP) Other (Q71GOTHR)***

**Question 7.2 ‐ Time (Q72TIME)**This question is intended to find out the number of hours children spent on household tasks in the last week.

**Question 7.3 ‐ Attending school (Q73ATTEND)** This question is for the interviewer to answer. Check if the respondent has answered “Yes” to Question 2.2. If the answer is “Yes”, go to Question 7.4, otherwise go to Question 8.0 which is the end of the interview.

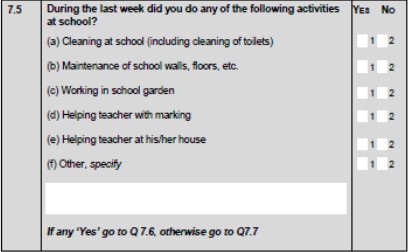
**Question 7.4** there should either be a 'Yes' or 'No' answer to all of them.



##### After school (Q741AFTERSCH)

***Before school (Q742BEFORESCH) Weekends (Q743WEEKEND)***

**Question 7.5** In order to be certain that the categories had been answered, there should either be a 'Yes' or 'No' answer to all of them. If there is any 'Yes' in Question 7.5, go to Question 7.6, otherwise skip to Question 7.7.

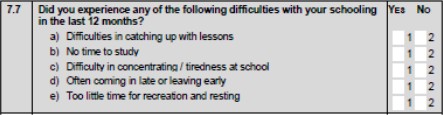


##### Cleaning at school (Q75ACLEAN) Maintenance (Q75BMAINTENANCE) Work in garden (Q75CGARDEN)

***Help teacher with marking (Q75DMARK) Help teacher at home (Q75EHOUSE) Other (Q75FOTHR)***

**Question 7.6 ‐ Time (Q76TIME)** This question is intended to find out the hours the child spends on the school activities mentioned in Question 7.5.

**Question 7.7**



##### Difficulty in catching up (Q77ADIF\_CTCHUP) No time to study (Q77BNOTIME)

***Difficulty in concentrating (Q77CCONCENTRATE) Punctuality (Q77DPUNCTUAL)***

##### Leisure time (Q77ELEISURE)

**Question 7.8**This question is asked for children who are attending school. (See Education code list).

##### Subject 1 (Q78ASUBCODE)

***Subject 2 (Q78BSUBCODE)***

Child Labour

Involvement in at least one form of child labour (Child\_Labour)

##### Final code list:

1 = Involved in child labour

0 = Not involved in child labour

## Data set 2

One line/record is one household.

#### Geography Type (Geo\_Type\_Code)

Classification according to the settlement characteristics

##### Final code list

1 = Urban

2 = Traditional

3 = Farms

#### Province (Province)

South African provinces as at December 2005 released by the Municipal Demarcation Board in January 2006.

##### Final code list

1 = Western Cape 2 = Eastern Cape 3 = Northern Cape 4 = Free State

5 = KwaZulu‐Natal 6 = North West

7 = Gauteng

8 = Mpumalanga

9 = Limpopo

#### Stratum

6‐digit number representing stratum formed during Master Sample 2006 where: digit 1 = province based on 2005 provincial boundaries;

digits 2 and 3 = metro/non‐metro; digit 4 = geography type.

***Valid range:*** 10101–90401

#### Metro/non‐metro (Metro\_code)

Derived variable: Derived from stratum

##### Final code list

01 = WC – Non‐metro

02 = WC – City of Cape Town 03 = EC – Non‐metro

04 = EC – Buffalo City

05 = EC – Nelson Mandela Bay 06 = NC – Non‐metro

07 = FS – Non‐metro 08 = FS – Mangaung 09 = KZN – Non‐metro 10 = KZN – eThekwini 11 = NW – Non‐metro 12 = GP – Non‐metro 13 = GP – Ekurhuleni

14 = GP – City of Johannesburg 15 = GP – City of Tshwane

16 = MP – Non‐metro 17 = LP – Non‐metro

## Dataset 3

One line/record is one child.

For the Q54TIME and Q72TIME questions the questions were asked in a slightly different way. Instead of asking how many hours per week they spent doing a certain activity, they were asked how many hours they spent on that activity each day of the week. So, their data was recorded by day instead of being added up for the whole week. This level of detail per day is included in dataset 3. You need to aggregate this back to the number of hours per week.