ASSIGNMENT
DATA PREPARATION AND CLASSIFICATION: CHILD LABOUR CASE

1 CONTEXT

In recent years, the Belgian government has taken an active interest in promoting the well-being of children worldwide. As part of this effort, they have funded programs aimed at assessing the quality of life of children across the globe.

Your team is working with the **"A Better Future" NGO**, a leading organization focused on **monitoring and combating child labour around the world**. Recently, the NGO conducted a **comprehensive survey of children between the ages of 7 and 17 in Africa**, with the goal of better understanding the extent and nature of child labour in the continent.

Your team's objective is to **prepare and analyse the data collected in this survey and develop a classification model that can accurately predict whether a child is involved in child labour or not**. This model will be a **powerful tool for identifying the key predictors of child labour and taking action to prevent it.** By gaining a deeper understanding of the factors that contribute to child labour, we can work to create a better future for children around the world.

2 DATA

The survey was conducted in **2022** and collected more than **13,000** responses.

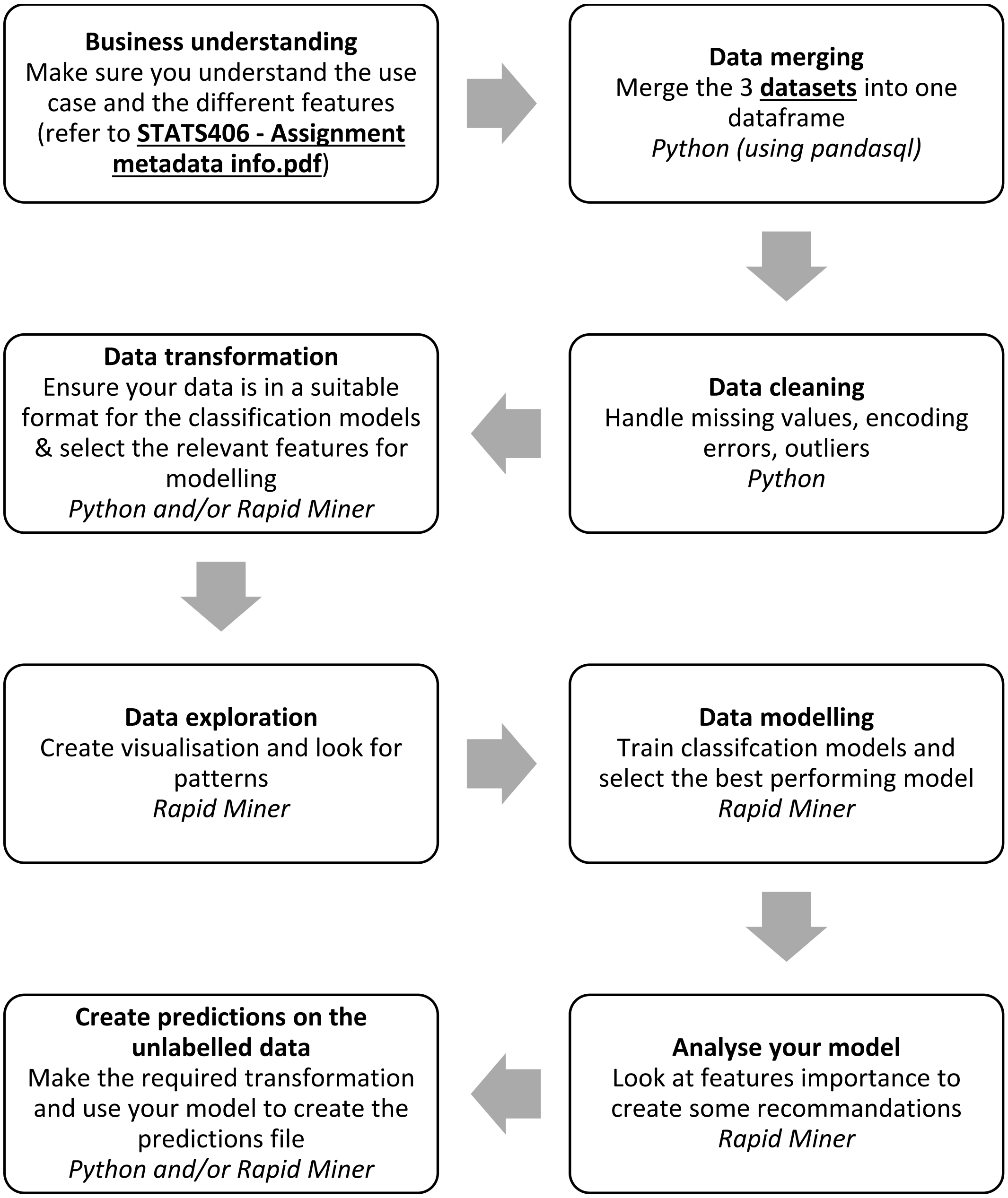
10% of the record has been kept away to evaluate the performance of your model on unseen data. You can find those records under **unlabelled_data.csv**.

You will be using the remaining 90% of the records to train and evaluate your models. Those records have been organized into three datasets, each of which is represented as a CSV file: **dataset1.csv**, **dataset2.csv**, **dataset3.csv.** These datasets include valuable information on various aspects of the children's lives, including their demographics information, educational background, …

All the details about the collected data can be found in the "**STATS406 - Assignment metadata info.pdf**" file.

3 WHAT IS EXPECTED FROM YOU?

You can find below a schema of the **general workflow you should follow for this assignment**. Note that even if this process is here depicted as a linear process, you should work in **iterations**. For example, you may go back to the data cleaning step after the data exploration step. You will also find in the schema the tool that we recommend for each step.



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

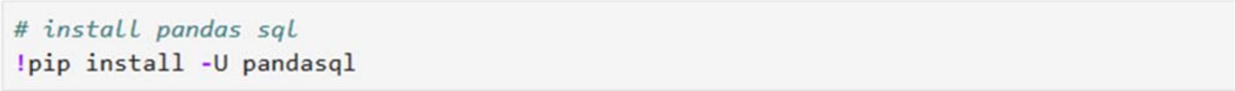
Data exploration
Create visualisation and look for patterns
Rapid Miner

Data modelling
Train classifcation models and select the best performing model
Rapid Miner

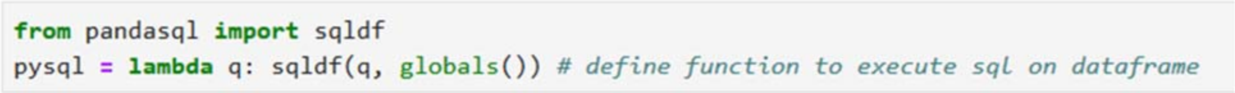
Analyse your model
Look at features importance to create some recommandations
Rapid Miner

Create predictions on the unlabelled data
Make the required transformation and use your model to create the predictions file
Python and/or Rapid Miner

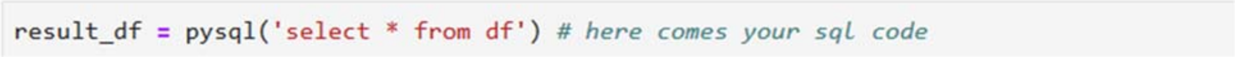
You will have to perform the **data merging and cleaning in python**. This means that you are not allowed to manually modify the input files. For this, we advise using pandasql. Pandasql will have access to all pandas dataframes you create (if you read in the csv files, but also if you create new dataframes) and will always return a pandas dataframe (that you can also use in next pysql query). You will have to run the following lines to initate pandasql:



```
# install pandas sql
!pip install -U pandasql
```

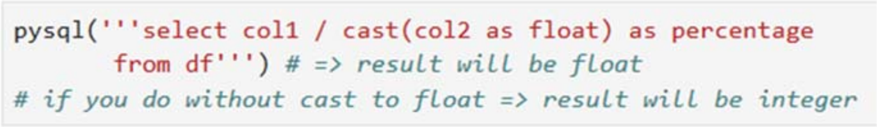


```
from pandasql import sqldf
pysql = lambda q: sqldf(q, globals()) # define function to execute sql on dataframe
```



```
result_df = pysql('select * from df') # here comes your sql code
```

Be aware that if you want to divide two integers by each other, for instance to compute a percentage, you will have to cast the integer in the denominator to a float:



```
pysql('''select col1 / cast(col2 as float) as percentage
       from df''') # => result will be float
# if you do without cast to float => result will be integer
```

In addition, pandasql is based on sqlite framework. If you need any specific sql function, such as how to extract a substring from a string, you can find some information here.

Finally, if you need to perform some date operations (such as transforming a string to a date and extracting a month number from it), we suggest using pandas functions as we have seen in session on Data Quality.

4 DELIVERABLES AND DEADLINES

You are expected to deliver:

1. A **report** (in PDF, max. 8 pages) containing at least:
    a. A **data preparation part** where you explain how you prepared and cleaned the data. This should include the following points:
        i. **Data merging**: explain how you merge the three datasets using python (we recommend using pandasql in python for this section).
        ii. **Data cleaning**: explain how you handled missing values, possible encoding errors and outliers, casting columns to the correct type, … (using python).
        iii. **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.
    b. 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)
    c. A **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.
    d. An **assessment part** where you explain how you evaluated the model.
    e. A **recommendation part** where you advise, based on your findings, the NGO about how to prevent child labor in South Africa.
    f. An **appendix section** (size free) containing at least:
        i. The schematics (a screenshot, not the actual model) of your process in RapidMiner.
        ii. A discussion of the most important design and analysis decision you had to make.
2. 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**.
3. The **Rapidminer process** (.rmp) used.
4. The **cleaned data frame in CSV format** that you used as input for the RapidMiner process.
5. 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 **<u>unlabelled_data.csv</u>**.
6. 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

[1] PS: you can download your Notebook as a .py file from the menu when it is running on Google Colab using File => Download .py

presentation that should be used during the presentation, without any possibility to modify it between the deadline and the presentation.

Failure to turn in one of the 6 files will result in a 0 for the whole assignment to all the members of the group. Please plan accordingly.

The report, python code, Rapidminer process, cleaned data in CSV and the predictions file (number 1 to 5 above) must be **uploaded on the UV no later than on the 24/04 11:59pm.** No delay, for any reason other than a generalized unavailability of the UV for at least 30 minutes during the last hour before the deadline, will be allowed.

For the presentation in PDF format (number 6 above), you must upload it on the UV **no later than on the 08/05 11:59pm.** No changes to the presentation will be allowed after this deadline.

On the 12/05 or 13/05, you will be **presenting your results**. The format will be as follows: **15 minutes of presentation and 10 minutes of Q&A**. **All members should be attending the final presentation**. A person not attending the final presentation will get 0 for the assignment. The only exception accepted is a valid medical certificate.

5 EVALUATION CRITERIA

This assignment will count for **40% of your final grade**, divided between:

- **Report: 20%**,
  - Data preparation and exploration: 10%
  - Modelling, assessment, and recommendation: 10%
- **Presentation: 20%**.

6 SOME ADVICE

Do not consider this assignment as a kind of “programming assignment”. When doing data science, the tentation is great to go directly to the technical matters and start crunching numbers. Don’t be fooled, though: this is a business endeavor and should be approached as such. Think of the kind of model you want to build, given the information you have, discuss it thoroughly, maybe test it first by building a mock presentation of your expected result and see if the way you present the problem makes sense.

Consider the time constraint and the goal here: more often than not, a perfectly good analysis is undermined by and underwhelming presentation and ends up being only an OK-ish project because the team invested too much in the analysis and not enough in the presentation. On the other hand, do not desist too fast and try to put lipstick on a pig by investing all your time in presenting mediocre results (we will likely see through it).

Before getting started on the Rapidminer model, you might want to look at Chapter 15 of the reference book “Data Science” from V. Kotu (a copy of the book is on the UV page of the class). It goes through a tutorial on how to use Rapidminer to perform the kind of analysis that is asked here.

7 SOME TIPS

Please find below some tips or useful tricks that will help you out in this assignment:

- If you can't read or write to your google drive as you can't find the right path, you can do
  ```
  import os
  os.listdir('/gdrive/')
  ```
  This will print out all the files and folders of the specified path, like this you can incrementally check if your path is correct by each time adding the next folder to the path in the listdir function
- If you need to combine two or multiple dataframes having the same columns into one dataframe, think about using the `UNION ALL` operator in SQL as we did in TP 2 here.
- In python, `df.colums` will print out the column names of your dataframe so that you can check if there are no spaces or other specific characters
- In python, `df.dtypes` will print out the types of each column, if it marks object, it is probably a string.
- You will notice that the dataset has missing values. You will need to understand how the questionnaire was made and answered (refer to the metadata.pdf file) to understand how to deal with those missing values. For example, Q16Mothparthh is skipped, so is missing, if the respondent answered ‘no’ in Q15Mothaliv.
- Make sure to use the “Child_Labor” as target variable in your modelling part.
- Since the target variable is characterized by quite an unbalanced distribution in both the training and test sets, we would suggest having a look at chapter 15.5 SAMPLING AND MISSING VALUE TOOLS (pages 513 to 516) in the Reference book (you can find here the link to the pdf version on the UV). There you can find useful suggestions on how to handle unbalanced datasets with Rapidminer.
- For the recommendation part, an idea could be to look at the variable's importance of your model. This could give you insights about which variables have a high impact on whether a child would be predicted to be involved in Child Labor or not. For example, for random forest models, you can refer to Weight by Tree Importance - RapidMiner Documentation.

8 WHAT TO DO IF YOU HAVE QUESTIONS/PROGRAMMING ERRORS?

Whenever you have a programming error, try to look first yourself by copying and pasting the error in google and look for answers (usually stackoverflow.com is a good place to find all your answers) 😉

If you are stuck and/or have more general questions about the assignment:

- For general questions/problems about the assignment, you must ask them in the **Teams Channel “Assignment”.** This to make sure everyone has the same information.
- For more private questions related to your group, you can **create a group chat via Teams with the three assistants** (Thomas Dekelver, Oliviero Gianfagna and Eléonore Charles) and send your message.

NOTE that we will **not answer questions** sent via **email** and/or sent to **only one of the assistants**.