UNIVERSITÉ LIBRE DE BRUXELLES ULB

FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

INFO-F-422 - Statistical foundations of machine learning
Project 2020-2021

Jacopo De Stefani, Théo Verhelst, Gianluca Bontempi

April 20, 2021

The project counts for 50% of your final grade (i.e. 10/20). This project has to be developed by a team of 3 students registered to the class. Any project returned by a team composed by a different number of students will not be considered. The project shall be completed independently and it shall represent the sole efforts of the team submitting the assignment. The result of another team efforts, or the copy of another team efforts (current, or past, semester(s)), is considered academic dishonesty and will be punished accordingly.

1 Goal

The goals of the project are:

- To participate to the *"Pump it Up: Data Mining the Water Table"* DrivenData competition by implementing and assessing different supervised learning algorithms and different methods of feature selection in the related classification task.
- to select among the learning and feature selection techniques the ones which appear to be the most accurate and use them for submitting to the DrivenData competition.
- to report your analyses and results as a Jupyter notebook.

2 DrivenData competition

The goal of the competition is to improve maintenance operations of water pumps and ensure that clean, potable water is available to communities across Tanzania. In order to achieve that, a smart understanding of which waterpoints will fail is required. Your objective is to build a predictive model which is able to correctly predict which pumps are functional, which need some repairs, and which don't work at all, using data from Taarifa and the Tanzanian Ministry of Water. The model has to be trained using the *Training set values*, *Training set labels* files available on the DrivenData platform (see Figure 1), it includes roughly 60000 labeled samples and 40 features. The students should then predict the *labels* for the samples included in the *Test set values*, and submit them to the platform following the provided *Submission format*. The students should register using its ULB/VUB netid as username and accept the rules of the competition (notably no hidden additional accounts).

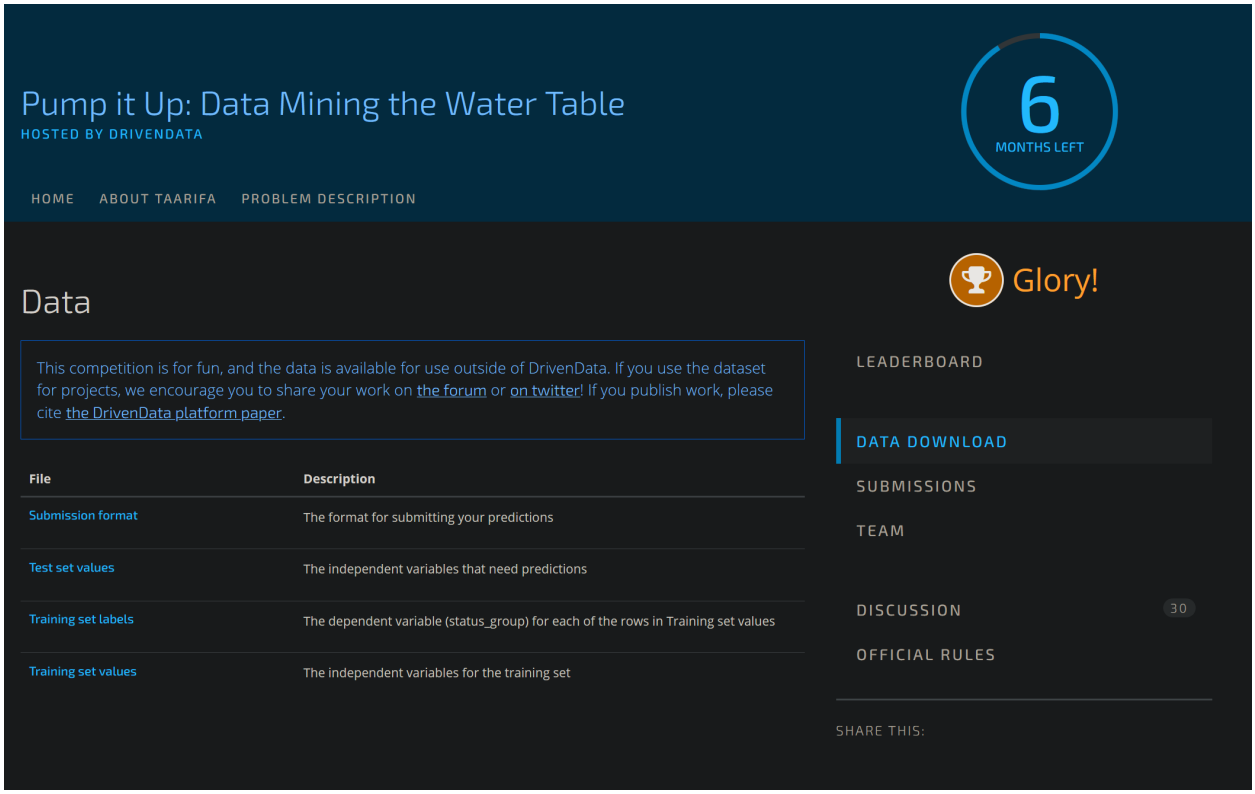

FIGURE 1 – Screenshot of the Data Download section of the DrivenData platform

3 The team

A project team has to be composed by exactly 3 (ULB or VUB) students registered to the class. Projects submitted by teams composed by a different number of students or by a student not officially registered will not be considered. Team composition must be submitted through a dedicated Google Form, available **here**, indicating the

1. the team name
2. the family name, first name and student ID of the three students,
3. link to the DrivenData team corresponding to the group

The team composition has to be finalized no later than **11PM of April the 27th 2021**.

4 Tasks

The team will have to:

1. implement in the R language a pipeline for data preprocessing, including missing value imputation, normalization (if required), feature engineering and feature selection. This procedure must be detailed in the notebook. The text must contain the list of relevant/selected variables and the motivation of their choice. The use of visualizations and tables to provide a better understanding of the data and the usage of formulas and pseudo-code to describe the feature selection procedure is strongly encouraged. Note one third of the score will be attributed on the basis of the quality of the documentation. **(3 points)**
2. implement in the R language a model selection procedure. This procedure must be detailed in the notebook and **exclusively use the packages listed in Section 5**. The text must mention the different

(and at least three) models (among those presented during the course) which have been taken into consideration and the procedure used for model assessment and selection. The use of figures, formulas, tables and pseudo-code to describe the model selection procedure is strongly encouraged. Note one third of the score will be attributed on the basis of the quality of the documentation.**(3 points)**

3. implement a learning procedure **using other R packages than the ones listed in Section 5** . This procedure must differ from the one in the previous point in terms of the classification model (e.g. a deep learning model, a gradient boosting tree) and/or the feature selection strategy. A procedure combining multiple models is also allowed, provided that it integrates at least a learner different from the ones presented in the practicals. This procedure must be detailed in the notebook. The text should justify the choice of this procedure, assess its accuracy with respect to the one developed in the point 2 and discuss the results. The use of figures, formulas, tables and pseudo-code to describe the combination of this novel procedure is strongly encouraged. Note one third of the score will be attributed on the basis of the quality of the documentation.**(3 points)**
4. On the basis of the procedure described in the previous steps the team must compute the predictions for the competition and submit them via the DrivenData website. The name of the team should appear in the official leaderboard of the competition. The link to enrol in the DataDriven competition is available **here**. **(1 point)**

5 Specifications

The team has to choose a learning method and a feature selection method among at least three alternatives. For the learning method (at point 2.), the only packages that may be used are those included in this list :

- stats/ridge (linear/ridge models)
- nnet (neural networks)
- tree/rpart (decision trees)
- randomForest (random forest)
- RSNNS (radial basis functions)
- lazy (nearest neighbours)
- e1071 (SVM)
- glmnet (LASSO/ElasticNet models)

For the point 3. the team is free to employ other learning methods, either already available online, or coded. The quality of the classification models during the selection process should be assessed by using classification accuracy. The report must be an R Jupyter notebook which has to specify and justify (with tables, figures) the selection procedures which led to the final choice. The team has to return, together with the report, the datasets employed in the notebook, the set of predictions submitted to the DrivenData competition and a video summarizing the whole predictive procedure.

6 Deliverables

The student team will deliver:

1. the implementation (in a R Jupyter notebook format) of the preprocessing pipeline, model selection and predictive procedure.
2. the datasets employed in the notebook (in .csv format).
3. the predictions submitted to the competition (in .csv format).

4. a video of max 10 minutes to present this project. The presentation should address the main points illustrated in the Jupyter notebook. Each of the three parts of the project must be presented by a different student of the team.

A template describing the structure of the project is available on the UV in the Project section.

Rules for project submission

To be read carefully!

1. The assignment should be made by teams of **exactly** three students. The team composition has to be finalized no later than **11PM of April the 27th 2021**.
2. The assignment will be graded on the implementation, the report and the video presentation.
3. The code should be **commented**.
4. The assignment will be handed in through the dedicated Homework module on the Virtual University.
5. All the deliverables will be put in a single archive, named `INFO-F-422_<STUDENT_ID>_<LAST_NAME>.zip` where `<STUDENT_ID>` and `<LAST_NAME>` should be replaced by the actual student id and last name of the student in the group.
   The archive should include:
   - **Python Jupyter Notebook** (`*.ipynb`)
   - **Report** (`*.pdf`)

   N.B. The report should contain a link to the video of presentation. Given the size of the videos and the format of the video, the video needs to be stored on a different platform than the Virtual University (e.g. Microsoft Stream, Youtube). In case of problems, get in touch with the assistant (`tverhels@ulb.ac.be`) to find an alternative solution.
6. Your project should be submitted on the UV no later than **11PM of May the 17th 2021**.
7. All the projects submitted after the deadline:
   - Penalized of one point if submitted **before 11PM of May the 18th 2021**.
8. Sharing of code is not allowed (you may, however, verbally discuss ideas on how to tackle the project).
9. This project counts for 50% of your grade (10 points). This project **shall be completed as a team and it shall represent your sole efforts**. The result or the copy of another team efforts (current, or past, semester(s)), is considered academic dishonesty. Plagiarism, in the sense of copy-pasting from existing reports or code is a serious issue.
10. Each project producing any error during its execution will receive a grade of 0/10.