Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
September Call, A.Y. 2020/2021

Last update: August 24, 2021

1 Project dates

Start date: August 24, 2021 at 23:59 AM CET
Due date: September 14, 2021 at 23:59 AM CET

Due date is a **strict deadline**.

2 Problem description

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. Since the virus spreads via these insects, the number of human infections is strictly related to their proliferation. One of the crucial factors conditioning mosquitos' life cycle is the weather.

In this project, you will predict the number of weekly dengue infections in two cities, namely San Juan (Porto Rico) and Iquitos (Peru). Specifically, you will build a regression model capable of inferring this number based on several meteorological descriptors.

2.1 Dataset

The dataset for this project contains weather measurements in tabular format. Each record is characterized by several numerical attributes referring to measurements sampled during a week. The following is a short description for each of them.

city: a short code that identifies either Iquitos or San Juan;
year: year in which the record was collected;
weekofyear: integer identifying the week within the year. The first week of the year is ‘0‘, the second one ‘1‘, and so on;
week_start_date: first day of the week;
max_temp_c: maximum temperature in Celsius degrees measured by the weather station;
min_temp_c: minimum temperature in Celsius degrees measured by the weather station;
avg_temp_c: mean temperature in Celsius degrees measured by the weather station;
precip_mm: precipitation in millimeters measured by the weather station;
diur_temp_rng_c: diurnal temperature range in Celsius degrees measured by the weather station;
PERSIANN_precip_mm: precipitation in millimeters measured by the PERSIANN satellite;[1]
NCEP_precip_mm: precipitation in millimeters measured by the NCEP Climate Forecast System Reanalysis (CFSR)[2] - all the following measures are collected by the same system;

[1] http://amir.eng.uci.edu/publications/19_HESS_PERSIANN.pdf
[2] https://www.ncei.noaa.gov/products/weather-climate-models/climate-forecast-system

NCEP_dew_point_temp_k: dew point temperature in Kelvin degrees measured by NCEP CFSR;
NCEP_air_temp_k: mean air temperature;
NCEP_humidity_percent: mean relative humidity;
NCEP_humidity_g_per_kg: mean specific humidity (g/kg). Specific humidity is the ratio of water vapor mass to total moist air parcel mass;
NCEP_precip_kg_per_m2: total precipitation (kg/m2);
NCEP_max_air_temp_k: maximum temperature in Kelvin degrees;
NCEP_min_air_temp_k: minimum temperature in Kelvin degrees;
NCEP_avg_temp_k: mean temperature in Kelvin degrees;
NCEP_diur_temp_rng_k: diurnal temperature range in Kelvin degrees.

The total number of infections during the week is reported on the feature named ***total_cases*** and is expressed as an integer $n \geq 0$.

The dataset is located at:
`https://dbdmg.polito.it/wordpress/wp-content/uploads/2021/08/student_files.zip`

Within the archive, you will find the following files:

- **development.tsv** (development set): a tab-separated values file containing the records from the development set. This portion does have the `total_cases` feature, which you should use to train and validate your models.
- **evaluation.tsv** (evaluation set): a tab-separated values file containing the records from the evaluation set. This portion does not have the `total_cases` feature.
- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a regression pipeline to predict the number of total cases of dengue fever within each week in the Evaluation set.

2.3 Evaluation metric

Your submissions will be evaluated through mean absolute error.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
10,2
123,0
21,23
345,12
42,3
...
```

The submission file must contains a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set, as an integer number.

> **Info:** The `Ids` in the submission file must correspond to the positions of the records in the Evaluation set. **The first record in the Evaluation set has Id=0, the second has Id=1 ans so on.**

- the `Predicted` label for the corresponding record.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `giuseppe.attanasio@polito.it`. Please refer to the guide on the course website, to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

> ◆ **Warning:** The report and the software have to be submitted by the due date reported in Section 1. This is a **strict deadline**.

Submission All the required files (i.e. for the report and the software) must be included in a **single *.zip*** file[3]. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section.
Please use as description: **report_exam_september_2021**.

Formatting rules The formatting rules for both the report and the software are described in the exam rules document. You can find it on the course website.

[3] A ZIP archive is a ZIP archive, not a RAR, a 7z or, a tarball archive, nor any of those renamed with a trailing `.zip` extension.