Data Science Lab: Process and methods
Politecnico di Torino

Project description
Winter First Call, A.Y. 2020/2021

Last update: January 10, 2021

1 Competition dates

> **Start date**: January 10, 2021 at 23:59 AM CET
> **Due date**: January 31, 2021 at 23:59 AM CET
>
> Due date is a **strict deadline**.

2 Problem description

This project consists in the prediction of a quality score associated with a wine review. Technically, you are required to build a regression model capable of inferring the wine's quality given the content of the review.

❖ **Warning:** For this competition, you will not be allowed to use external datasets other than that provided for the competition. Adoption of external resources will result in failure of the exam.

2.1 Dataset

The dataset for this project contains wine reviews in tabular format. It counts 150.930 entries, each of which referes to a review expressed by an expert on a specific wine.

Each review is characterized by both numerical and categorical attributes. A textual description is provided as well. The quality score is indeed reported on the feature named `quality`. The dataset is located at:

`https://dbdmg.polito.it/wordpress/wp-content/uploads/2021/01/competition_dataset.zip`

Within the archive, you will find the following files:

- **dev.tsv** (development set): a tab-separated values file containing the reviews from the development set. This portion does have the `quality` feature, which you should use to train and validate your models.

- **eval.tsv** (evaluation set): a tab-separated values file containing the reviews from the evaluation set. This portion does not have the `quality` feature.

- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a regression pipeline to assign a quality score to each record in the Evaluation set. The score expresses a overall evaluation assigned to the wine.

2.3 Evaluation metric

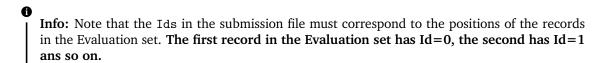Your submissions will be evaluated through r2_score.

3 Submit your result

Submission file   In order to get your results evaluated, you have to upload a result file on our submission competition platform. The submission file has to be a `.csv` file formatted as follow:

```
Id,Predicted
10,10
123,11.3
21,54
345,23
42,35.823
...
```

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.

  > **ℹ Info:** Note that 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 on the competition material.

Submission platform   The submission platform is the same you used during the course laboratories. Therefore, you have to use your personal key[1]. Please refer to the guide on the course website, to go through the submission procedure.
You can find the competition 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. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section.
Please use as description: **report_exam_winter_2021_first_call**.

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.

---

[1]You should have received a new key recently: check your email inbox. If that is not the case, please send an email to giuseppe.attanasio@polito.it