Big Data Computing Final Project Guidelines

Gabriele Tolomei

Sapienza University of Rome, Italy

Email: tolomei@di.uniroma1.it **Homepage:** https://www.di.uniroma1.it/∼tolomei

Scope of the Document

This document provides students of the **2021-22 Big Data Computing** class (hereinafter, "you") with a list of *guidelines* for developing their projects, which are mandatory for final grading. Please, read through the *whole* document carefully.

1 Project Proposal

No matter what project you decide to work on (more on this in the sections below), **this must first be approved**. In order to get such approval, you have to come up with a written project proposal, i.e., a half-page PDF document, that contains at least the following information:

- The problem/task you are planning to address (e.g., binary classification);
- The dataset(s) you will be using along with their references, or how do you plan to collect data if no dataset is already available for achieving your goal;
- The methods you would like to experiment with (e.g., SVM, logistic regression, random forest, etc.);
- The evaluation framework you will use to assess the quality of each method (e.g., accuracy, precision-recall, AUROC, etc.)

Project proposals are **mandatory** and must be submitted via Moodle using the corresponding assignment "*Project Proposal Submission Week*" that you will find within each exam session.

IMPORTANT NOTE: From this year, project proposals must be submitted within a specific time window (please, refer to the example below). This is to avoid keeping track of multiple proposals that arrive "in dribs and drabs", scattered across several different times of the year.

To be more specific, consider the following concrete example. Suppose that the project submission deadline that you are aiming for is set for **June 27**. Although it is hard to fix the right amount of time that you will need to develop your project (as this, of course, depends on several aspects), a good rule of thumb is to reserve **at least one month** of work. This means that your proposal **must be approved** at least one month before the project submission deadline that you are targeting. Furthermore, to allow me considering your proposal, give you some feedback, and eventually send my definitive approval on time, I must receive it slightly before. Getting back to the example above, you should expect to issue your project proposal within the submission week **from May 21 to May 27**.

Remember: Do not start working on a project if you didn't get it approved first! In any examination session, if you submit a project on time but its proposal hasn't been previously approved, this will **not** be accepted nor evaluated for grading.

2 Project Requirements

Projects must of course refer to a typical big data task, such as those seen during classes: e.g., information retrieval, clustering, regression/classification, recommendation, graph analysis, using large datasets in *any* application domain of interest.
The development environment must be the one of those used throughout the course, namely PySpark in combination with either Google Colab or Databricks Community Edition platform.
Projects can be done either **individually** or in group of **at most 2 students**, and they should be accompanied by a brief presentation written in english (e.g., a few PowerPoint slides). In the latter case, I am expecting a more complex project so as to justify the effort of two people.
There is no restriction on the technique(s) you should use to solve your task (e.g., K-means, logistic regression, matrix factorization, artificial neural networks, etc.) In other words, you can experiment with any technique you want – providing it is relevant to your goal and hopefully comparing more than one with each other – including also those not covered in class. In fact, you are strongly encouraged to explore non-standard methods!

3 Project Selection

There exist several resources that contain many project ideas as well as the related datasets to work on; among those, I would recommend the following ones:

- Kaggle
- UCI Machine Learning Repository
- Awesome Public Datasets

Alternatively, you can also come up with your own project idea, as long as it satisfies the requirements indicated in the section above. This is particularly encouraged and appreciated, as recently I have started reviewing very similar projects... 😒

4 Dataset Policy

It is **strictly forbidden** to use datasets which are downloaded without the required permissions and/or coming from illegitimate sources. You can, of course, build your own dataset to work with but that must be collected according to the law, and all the steps performed must be properly documented and acknowledged.

5 Project Evaluation

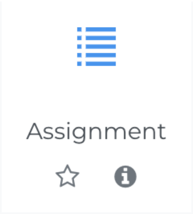
Projects will be evaluated according to the following **5 criteria**:

- "**Big-Data-ness**": This criterion will check if the "big data" requirements of the project are actually met. It will provide answers to questions like "*How complex the task is?*" or "*How large is the dataset used?*".
- **Coding**: This criterion will evaluate the style of programming. It will provide answers to questions like "*Is the source code easy to read, reusable, and well documented?*" or "*How much effort would it take for the source code to get into a hypothetical production system?*".

- **Methodology**: This criterion will assess the soundness of the methodological approach used to achieve the project's goal. It will provide answers to questions like "*Does the pipeline implement all the stages needed?*" or "*How reliable are the results obtained?*".
- **Originality/Impact**: This criterion will validate both the novelty of the task that the project aims to solve and the creativity of the solutions implemented. It will provide answers to questions like "*Is the task novel/impactful or has it been already deeply investigated?*" or "*How conventional are the solutions adopted to solve the task?*".
- **Presentation**: This criterion will judge the quality of the presentation attached to the project. It will provide answers to questions like "*Is the project well defined and its goal clearly stated?*" or "*What is the level of understanding of the challenges associated with the specific task?*".

6 Project Submission

Projects will be submitted for grading to the Moodle web page of the course. More specifically, for each examination session there will be an *Assignment* entry on Moodle, which will allow you to upload your project material (along with the corresponding project proposal that you are expected to send before). Moodle assignments are identified by the following icon:



Assignment

☆ ℹ

For example, suppose you are ready to submit your project on the June 2022 session. On the Moodle web page of the course, you will see an entry called "*June 2022 Exam Session*", along with an assignment corresponding to that session named "*June 2022 Project Submission Week*". This assignment has a **one-week** time window, within which you must submit your project (e.g., from **June, 21 at 00:00** to **June, 27 at 23:59**[1]). More generally, everyone who wants to submit their project **must** upload the material within the deadline established by the project submission week of a specific examination session. Note that within the same Moodle entry "*June 2022 Exam Session*" you will find another assignment called "*Project Proposal Submission Week*", where you must have previously submitted your proposal by the specified deadline (e.g., from **May, 21 at 00:00** to **May, 27 at 23:59**).
The project material **must** be packaged inside a **single archive file**, namely a (compressed) folder (e.g., `.tar`, `.tgz`, `.bz2`, etc.) containing the following **two items**:

- A notebook file (`.ipynb`) with all the source code of the project;
- A presentation (e.g., PowerPoint or PDF slides) with a description of the project, the main choices made to accomplish the task, and the results obtained.

Please, consider that the notebook must be "ready-to-execute": in other words, it must contain everything to be run properly and successfully (e.g., environment setup, library dependencies, etc.)

[1] Central European Time (CET) or Central European Summer Time (CEST), if not otherwise specified.

To ease the grading process, it may be helpful to setup a naming convention for the project submission. The uploaded folder should be named as follows: `X(_Y).Z`, where:

- `X` and `Y` are *student IDs* of the project team;
- `Z` = Archive extension (e.g., `tar`, `tgz`, `bz2`, etc.)

In the case of a single-person team, the folder will be named as `X.Z`, whilst in the case of two-person team, `X` and `Y` will be the ID of the first and second student, *after they have been alphabetically ordered by their last name.*
For example, if the project is done by a single student whose ID is "12345", then `X`=`12345` and the folder will be named `12345.tgz` (or similar extension). Instead, if the project is done by two students: *Clark Kent* whose student ID is "67890" and *Bruce Wayne* whose student ID is "12345", then the project folder will be named `67890_12345.tgz` (or similar extension). In addition, files within the project folder (i.e., the notebook and the presentation) should follow the same naming convention.

NOTE: Teamwork projects must be submitted **only once by one member** of the team, who is the responsible for the submission.

7 Project Discussion

Right after the project submission deadline of an examination session and before the next one, there will be an oral discussion session.
The oral session is composed of **two parts**:

- An *oral presentation* (supported by few slides): you will describe the main goal of your project, a comparison of the proposed approaches, and the main results obtained (**max. 20 minutes**[2]);
- A *project demo*: you may be asked to show a working demo of your project by running (part of) your notebook, and to motivate the choices you made.

The entire session will be in english. Questions about any other topic addressed during the course may also be asked, but those can be answered either in english or in italian, as you prefer.
The oral session is public, and therefore *everyone* is welcome to join it!

NOTE: Unless otherwise specified, oral discussions will take place in person. Exceptions can be considered on a case by case basis, allowing candidates to participate also from remote via Google Meet or Zoom. Anyway, you will be notified in advance on how to attend the examination session.

8 Honor Code

The basic principle under which you are expected to operate is that you should submit your own work. Of course, if you are part of a two-member project team this principle will naturally extend to the team as a whole. More specifically, attempting to take credit for someone else's work by turning it in as your own constitutes *plagiarism*, which is a serious violation of basic academic standards.
To observe the honor code, you are invited to follow the rules below:

[2] In case of two-person teams, this will be extended to 30 minutes and each member of the team must be actively involved in the oral presentation.

- You must not submit or look at solutions or program code that are not your own;
- You must not share your solution code with other students, nor ask others to share their solutions with you;
- You must indicate on your submission any assistance you received.

Please, be aware that all submissions are subject to automated plagiarism detection.
As a final remark, many forms of collaboration – if legitimate and properly acknowledged – are acceptable and indeed encouraged.

Good luck and unleash your creativity!