Programming for Data Science
Final Project

niccolo.marastoni@univr.it

December 2022

The project consists of an analysis performed on a dataset chosen by each student. The dataset needs to have a decent quantity of data points (rows), and each data point needs to have some features (columns). Once an appropriate dataset has been found, it has to be submitted to the professor along with an outline of the ideas for the project. **Wait for it to be approved before starting to work on it**.

kha khá số lượng của điểm dữ liệu

dữ liệu thích hợp

What follows is a rough outline of how a project can be structured and evaluated. These steps serve as a guideline for anyone that appreciates a more structure approach to the exam. You are free to take these steps in any order and focus on any of them as you see fit in your project.

1 Data Exploration + Data Wrangling

Completing this part without any issues guarantees passing the class.

This part is worth: **20 points**.

1. Explore the dataset (6 points): Your first task is to explore the data as seen in class, finding correlations between attributes and finding some interesting aspects that justify the next parts of the analysis.

2. Clean up the dataset (6 points) This step encompasses the replacement of null values with appropriate data or their outright removal. During this step you are encouraged to modify the data according to the considerations matured during the exploration.

3. Show some interesting plots (6 points) An essential skill of a data scientist is being able to show the important information by using easily understandable graphs. Use the libraries introduced in class to showcase some interesting aspects of the dataset.

An additional **2 points** can be awarded for code cleanliness (hence the total being 20 points instead of 18).

2 Find a model that explains the data

For this part you can use any tool you are familiar with, be it statistical methods, machine learning or even deep learning.

Possible ideas:

- regression model if you are trying to predict continuous values (see stock market example in Lecture 13)
- classification model when you have discrete and finite labels (see Titanic model in Lecture 14)
- clustering if you want to find underlying structures of the data, without prior information

This part is worth: **4 points**.

3 Build a presentation

This ties in nicely with the 3rd step in part 1: you need to be able to present the data in an intuitive way. A nice way to simplify the presentation is removing the underlying code from the view. In order to get full marks in this part you need to submit a presentation built with streamlit (as seen in class) or similar libraries (Flask, Django etc.).

một cách trực quan

The presentation needs to showcase all the parts outlined in the project so far. This part is worth: **4 points**.

4 Track progress through Git

You should create a public GitHub repository for your project, or you can create a private repository and change it to public when you are ready to submit.

Using Git as a showcase for your portfolio is only a small part of the project, it is important that you also show the progress of your project in the commit history.

This part is worth: **2 points**.

5 Project Submission

Once your project is finished, you can send me the link to your public repository on GitHub via email. You can also send your project in a .zip attachment if you decided to forgo the GitHub part (try to avoid this) and make sure to **include your colab notebook**. The email subject will have the following structure: ”<name> <surname> - <studentID> - Final Project Submission”.

For example: ”Niccolò Marastoni - VR364254 - Final Project Submission”. Any submission that does not adhere to this structure will be discarded.

Exam: The project submission alone will not grant you a grade in this part of the class. In order to successfully complete the exam you will need to send a project that nets you **at least** 18 points. After this, you will need to take a short oral exam with the professor on the scheduled dates that you can find on esse3. The exam will be **in english** and will mainly test your knowledge on the project.

Make sure to send your project at least one week before the exam date.

Plagiarism will not be tolerated. This project is meant to showcase your ability to use the programming concepts explored during the class, so you have to submit **your own code**.
PS. StackOverflow does not count as plagiarism.

Exam date: The exam is scheduled for February 9th, 2023. After you send your project you will be assigned a time for your oral exam. The exam will be held between 8:00 and 18:00 in classroom 1.02 in Ca' Vignal 3. You will need to bring a laptop with which you can showcase the project.

Further questions: If you find that certain things about the project are not made clear in this document or if you have any doubts, please send me an email and I will get back to you as soon as possible.