Long Assignment 2025/2026

Alipio Jorge, Nuno Guimarães

November 1 (version 01.11.2025)

Objectives

Before a food product goes to market, panels of hired tasters give their opinions in order to help forecast the market acceptance. The composition of the food is important and tasters give their scores about different aspects of appreciation. These are called sensory tests. Other external factors related to brand and marketing are important to product success as well.

Is it possible to reduce or at least assess the risks of having a new product in the market? Is it possible to dispense or reduce sensory testing?

Dataset

The data used in this assignment is dataset inspired in the activity of a company from the Porto area.

The data is provided as an excel file and contains descriptions of products, tastings and external factors and has two potential target variables. They are `overall_appreciation` and `success`. The first one is a summary of the taster's opinion and the latter says whether the product was successful in the market or not.

The data dictionary can be also found in the excel file.

Note that the data may have imperfections.

You will need to create the train and test datasets:

1. Use the first 1600 rows for training or cross validation.
2. Use the next 200 rows for validation.
3. Use the last 200 rows for testing.

Evaluate the stratification of the data for the different classes. While this may not be strictly necessary, it is one possible approach to use.

Note that the dataset will need proper pre-processing.

Guidelines

This data science problem should be approached by following the CRISP-DM methodology (http://jbusse.de/2019_ws_dsci/crisp-dm_phases-tasks-outputs.html). You have to understand the business problem, propose success criteria and see how it can be translated into machine learning problems. Then you look at the characteristics of the data and you perform the required explorations, visualizations and transformations. Next step is to identify insights, develop predictive models and to evaluate them in order to validate if they are helpful in the business problems. During the whole process take notes, always identify the questions you want to answer and think before you act: "why is this plot or this transformation useful". You can perform some operations just for the sake of training but you should be aware of that.

The result is a **report** in the form of a **notebook** with clear explanatory text and code that works showing results. The report should be clear, as concise as possible and it should be easy to read and to follow. You will be telling the story of your approach to this problem, so it should have a good narrative flow. Always explain what you are doing, why you are doing it, what are the results and what do you take from those results.

Always think of the business problem and of future operational conditions that may limit applicability of the model. Not everything that we can do with a dataset is doable under operational conditions.

Suggested structure

A report containing:

1. Business understanding
   - Give your view of the business problem following the CRISP-DM list of outputs when adequate.
2. Data Understanding
   - Looking at the raw data, describe variables according to their types: interval-scaled, binary, nominal, ordinal, ratio-scaled. Be aware that there are specific methods suitable to each type of variable.
   - Perform a preliminary analysis (summaries, spread measures, histograms, boxplots, density). These are interesting to be applied to the raw data to "uncover'' inconsistencies, outliers, duplicates etc.
   - Perform bivariate analysis (correlations, regression)
   - Provide any insights about the data and the problem that you may have found.
3. Data Preparation
   - List of main changes that can need to be performed to the raw data, including feature selection.
   - Describe the potentially useful ones and their results in terms of data.
4. Modeling: consider the balanced and the non-balanced versions of the dataset as 2 separate problems. Firdt work with the balanced data and then with the non-balanced data. Try each of the methods below, select hyper parameters using default values and empirical analysis. Separate a test set and use cross-validation on the rest of the examples. Visualize models when possible, visualize results, produce aggregating tables with good insightful summaries of the results, and whatever other tools you may find useful.
   - Nearest neighbor
   - Bayesian Classifier
   - Decision Trees
   - Tree ensembles
   - Support Vector Machines
   - Neural Network Classifier
   - Comparison
5. Evaluation and Main Conclusions
   - What is the best model and the recommended data science procedure for each business problem?
   - What do you think that the business can gain from your data science effort?
   - What are the lessons learnt?
   - What is your summary of the achieved results?

To submit:

- a fully operational Jupyter notebook with the selected experiments as clear and concise as possible. Avoid output dumps. Recall that the report is going to be evaluated by your very busy professors and that they may have to skip many pages if your report is too long. Always highlight your best results. Please note:
  - The objectives for each experiment and plot should be clear so that the reader understands why it is worth to read a particular part.
  - The conclusion should be a short high level account of what was observed.

  - It is **not necessary to describe the methods** (unless requested, but you should know their concepts and how they work). It is more important to point out the differences in the methods and the reasons for the results in terms of methods characteristics.
- The project slides presentation.

Evaluation

- This assignment is worth the values described in sigarra, according to the course you are following.
- Components
  - Report 30%
    - Narrative 10%
    - Writing style 10%
    - Presentation 10%
  - Technical 70%
    - Pre-processing 10%
    - Diversity of the results for the experiments 20%
    - Correctness 30%
    - Conclusions 10%

Groups

Assignments are submitted by groups of 3 students. Different elements may have different grades. Other group sizes will not be considered. If different elements of the group have different levels of effort that should be stated upfrint by the group in the report.

It is advisable that the students from the same group perform overlapping work and only after that, exchange ideas with each other. Group work is important for learning from other people.

Submissions

Formal final deadline is **December 22nd 2023**, to be submitted in moodle, and only in moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

- Checkpoints:
  - In the class of the **27th November** there will be a checkpoint. Each group should present on desk an update with the status of the project. You will have around 5 minutes for this presentation, where you can show your current results and list your main difficulties.
  - Final presentations of the project on **December 18th** morning. All members of the group should be present.

Use of resources including Generative AI (pedagogical advise)

You, students, are learning how to do data science and the main aim of this exercise is that you gain skills. Skills are gained with different depth and robustness depending on how we use the learning resources. Effort must always be there. Beware of too easy routes - they usually teach us nothing. No pain, no gain.

Your **resources** are class materials, books, youtube videos, materials on the web, etc. (CM), programming documentation, such as the excellent manuals of pandas and sklearn (PD), search engines, e.g. google (SE) and generative AI, e.g. chatgpt or gemini (GAI).

Our **advise** is that you approach your problems by using these resources by this order and with different types of usage. When thinking of concepts, use CM, when you need to know how to use a function or what to use in a particular setting, use PD. SE are very important for more directed searches, looking for code

snippets, looking for more materials, finding fruitful discussions. Finally GAI gives us more when used for concept clarification or to find the starting point for a search or quest. Many other usages of GAI are productive but they should always have you leading, not the other way around.

An **example of good usage of GAI** is "Which hyperparameters are important for using neural networks and why?". A **bad usage** is "produce the code for applying neural networks for this problem". You can gain skills with the first type of usage whereas with the second you rely entirely on the skills of generative AI. That implies that any future employee will rather buy a license of chatGPT than hiring you (because it is cheaper). Moreover, when you gain skills you are ready to gain more skills and you are a valuable asset that can check what gen AI is producing. Having skills makes you think longer, deeper and depend less on externalities. Deep skills make you more robust and more resilient. If your skills are easy to get then anyone can do it.

Ethical principles

When submitting, students commit themselves to follow strong ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole of the work. All the materials used and consulted must be credited in the work.