



Machine Learning I - Group Project

Nata Visionaries

Bachelor in Data Science, 2025-26

17/Oct/2025, v.1.0.0

1 Introduction

Nata Visionaries is an ancient, not-for-profit brotherhood formed around Portuguese Custard Tarts, the **Pastel de Nata**. It is a charming organization, with a positive attitude towards life and society, with a significant obsession with the Pastel de Nata.

Bound by pastry and purpose, the Brotherhood is devoted to the ancient art of the Pastel de Nata.

Through careful tasting, endless testing, and the occasional sugar-induced revelation, members seek the perfect union of crisp and cream. Eggs over egos.

We are the alchemists seeking the pure gold of a perfect Pastel de Nata. Our mission is humble yet noble: to rule the palate, not the planet.

May every bite bring us closer to enlightenment... or at least to another Pastel de Nata.

Nata Visionaries heard about **Data Visionaries** and decided to hire your team!

2 The project

2.1 Background

Nata Visionaries are very methodical and organized, and have collected data from traditional bakeries in Porto and Lisbon to uncover what makes an exceptional Pastel de Nata. The information collected includes the recipe with quantities, information about the preparation, and information about the baking process. For each recipe, the resulting Pastel de Nata is classified as either "OK", meaning

that it deserves the seal of approval of the Nata Visionaries, or "KO", meaning that yes, you can still eat it, but it's not the real thing.

This classification is obtained by a process that is extremely accurate, but that is kept even more secret than the recipe for Pastéis de Belém. The only thing we know is that it involves destroying the Pastel de Nata without actually eating it. And the brotherhood has had enough of that nonsense!

2.2 Objectives

Your team was hired to build a model that can predict the quality based on ~~existing Nata~~ existing data. Just with the information about a certain production, your model will classify the Pastel de Nata as either "OK" or "KO".

The model will save thousands of Pastéis de Nata every year from the fate of "destruction without consumption". It will also provide information that can be used by different bakeries to improve on their recipes and production processes.

Your team is free to develop the work in whichever way you feel is preferable. It is nevertheless suggested that you use these steps:

- Business Understanding: *In crust we trust.*
As most business info is secret, the only information provided by the Nata Visionaries was that of the target variable and the metric that you should use;
- Data understanding: *The path to wisdom is dusted with cinnamon.*
Explore data to understand the main characteristics and limitations of the dataset;
- Data Preparation: *May our crusts be crisp and our spirits flaky.*
Building on the data exploration and insights, prepare data for modelling.
This may include the need for dropping existing features or for creating new ones.
- Modelling: *Our philosophy is simple: bake it until you make it.*
Experiment with different models that can predict the class of a Pastel de Nata; assess those models, using one or more configurations, to identify the top-performing models;
- Evaluation: *We don't sugar-coat the truth — we caramelize it.*
Evaluate the selected model on different metrics. Attempt further optimization.
- Deployment: *From Portugal with crust.*
Use this final model to predict the results for the test data for which you have no labels. Submit it to Kaggle.

3 Nata Data provided

3.1 Files provided

Three data files are provided, all in CSV format:

- **learn.csv** has the measures performed and the ground truth (quality_class);

- **predict.csv** has a structure similar to learn.csv, except that it does not include the ground truth. You will need to make a prediction for each record in this file, and submit it for evaluation (more on that later);
- **sampred.csv** is provided as an example of the structure of a predictions file. You could and can, submit this file. Just don't expect anyone to be impressed with the result.

3.2 Features

The dataset contains the following features collected for each Pastel de Nata:

| Feature | Type | Description | Unit |
|-------------------|-------------|---------------------------------|----------|
| origin | Categorical | Bakery origin (Porto or Lisboa) | - |
| preheating_time | Numeric | Oven preheating duration | minutes |
| baking_duration | Numeric | Total baking time at high temp | minutes |
| cooling_period | Numeric | Resting time after baking | minutes |
| sugar_content | Numeric | Sugar in custard filling | g/100g |
| salt_ratio | Numeric | Salt content in puff pastry | g/kg |
| egg_temperature | Numeric | Temp of eggs when added | °C |
| oven_temperature | Numeric | Final baking temperature | °C |
| cream_fat_content | Numeric | Fat percentage in cream | % |
| lemon_zest_ph | Numeric | pH of lemon zest infusion | pH scale |
| vanilla_extract | Numeric | Vanilla extract concentration | ml/L |
| egg_yolk_count | Numeric | Number of egg yolks per batch | count |
| quality_class | Categorical | Quality category (target) | OK/KO |

Table 1: Dataset Features

3.3 Categories

The federation classifies Pastéis de Nata in two quality categories. That information is stored in quality_class:

- **OK:** Seal of approval from Nata Visionaries
- **KO:** It's safe to eat, probably, but why would you?

4 Deliverables

Each group must submit one zip file with all the notebooks. Only one element of the group should submit that file. The notebooks included should be run from start to finish and include all the outputs. If Hyperparameter tuning takes too long, you can comment just that part before running the notebook in which it is implemented. Make sure the notebooks run properly in VS Code, as that'll be the IDE used to evaluate them.

Follow the below naming instructions:

The zip file must be named MLXX_Notebooks.zip (where XX is your group number with two digits). As an example: ML03_Notebooks.zip

Each file in the ZIP must be named MLXX_NBY_TITLE.ipynb (XX is your group number with two digits, Y is the number of the notebook and TITLE is an optional title.

The suggested package should include 5 notebooks (example names for group 03):

- **Notebook 1**, Data Exploration (ML03_NB1_DATAEXPLORATION.ipynb)
- **Notebook 2**, Preprocessing (ML03_NB2_PREPROCESSING.ipynb)
- **Notebook 3**, Feature Management (ML03_NB3_FEATUREMGMT.ipynb)
- **Notebook 4**, Modelling and tuning (ML03_NB4_MODELLING.ipynb)
- **Notebook 9** (yes, 9!), Complete final model and Kaggle predictions (ML03_NB9_FINAL.ipynb)

Both Notebooks 1 and 9 must be runnable directly from the initial data files, without going through the other notebooks to generate data files. In the case of Notebook 9, that's up to the point where the predictions file is exported.

For Notebooks 2, 3 and 4, it is acceptable that they must be executed in order (it may be the case that notebook 2 generates files used by notebook 3 and 4). You may opt to merge some (or all) notebooks 2, 3 and 4. In that case, make it very clear what it is that you're doing at each step of the process. Adjust the name of the notebook accordingly (ex: ML03_NB2_PREP_FEAT_MODEL.ipynb).

5 Kaggle Competition

As part of the journey, you will embark on a Kaggle (www.kaggle.com) private competition against the other groups.

As soon as you have a working model, you can make predictions for the predict.csv file, then submit them on Kaggle. We encourage you to do that as soon as possible, even if you're not happy with those first models. A few moments after your submission, the site will show the public score for your submission. This is the score obtained on ca 30% of the predict.csv file. This score will be visible for everyone to see, on the leaderboard.

The final score will be calculated, after the competition closes, based on the other 70% of the predict.csv file. This score will have a small contribution to your final grade on the project!

Keep in mind that the metric used for evaluation will be Accuracy.

6 Important Dates

| Milestone | Deadline | Penalty |
|-----------------------|----------------|--|
| Kaggle comp. closes | 17/Dec 23:59 | N/A: Late delivery not possible |
| Notebooks submission | 20/Dec 23:59 | 2 points per day or part, up to 3 days |
| Project Presentations | Day after exam | No show = 0 in project |

After Kaggle's deadline, you have three days for notebook delivery. You should not use them to make changes to the models, as that will create inconsistencies. You may use it to improve the quality of your code and to improve the overall contents, as there is no separate report.

The final scoreboard will only be made available after the notebooks' delivery deadline.

7 Project Evaluation

7.1 Grading:

| Item | Grade, scale 0-20 |
|---------------------------|-------------------|
| Data Exploration (NB1) | 4/20 |
| Preprocessing (NB2) | 2/20 |
| Feature Engineering (NB3) | 4/20 |
| Modeling (NB4) | 5/20 |
| Final model and HPT (NB9) | 4/20 |
| KAGGLE competition score | 1/20 |

7.2 Notebooks' evaluation

As there is no separate report, your notebook should provide all the contents that you would place in the report, but in markdown cells. For each notebook and for each section, explain what you're doing, then write the code, then interpret the results and mention insights you obtained. Notebooks will be evaluated on:

- **Documentation** - Is your reasoning correct? Is it well explained?
- **Code documentation** - Is your code functioning? Organized? Commented?
- **Choice of adequate techniques** - Are you using all relevant techniques, or are you missing something important? Did you add something that makes no sense?
- **Correct implementation** - No Errors, no inconsistencies, no data leakage.

NOTE! You will lose points if you don't follow the instructions in detail: only one submission per group, following the naming conventions. If you already submitted the notebook and identify some correction that needs to be made, you can do that until the deadline, without any penalization.

7.3 Final considerations

Remember that it is a group project but **your grade is individual**. During the presentation, you must be able to explain any part of what you did, actually show it in the notebook, and to make changes that may be requested.