8TH the last edition
The financial challenge of the year
NOVARTIS DATATHON
online

CHALLENGE AND SUBMISSION INSTRUCTIONS

Datathon 2025

Organized by Novartis Digital Finance Hub

November 2025

Contents

The Generics Problem

In the pharmaceutical industry, when a company develops a new drug, it is granted a patent that provides exclusive rights to produce and commercialize the product for a specific period. Once the patent expires, the drug reaches its *Loss of Exclusivity (LOE)* date. From that moment onward, other manufacturers are allowed to produce generic versions of the drug.

A generic drug is defined as a product that is equivalent to the brand-name medication in terms of dosage form, strength, route of administration, quality, performance, and intended use. Although generic products may contain different inactive ingredients (e.g., fillers or binders), these differences do not affect the therapeutic effect of the medication.

Generic manufacturers benefit from significantly lower development costs, as they are not required to repeat extensive clinical research. Instead, they must demonstrate *bioequivalence* to the original product by comparing pharmacokinetic properties such as absorption, distribution, metabolism, and elimination.

The entry of generics into the market typically leads to several consequences:

1. **Increased competition:** Multiple manufacturers producing the same medication increases market competition, which generally drives prices downward.

2. **Improved affordability:** Generic versions are usually sold at lower prices, making treatments more accessible for patients and healthcare systems.

3. **Greater access to medication:** Reduced prices expand access to treatment for a larger population, which may improve disease management outcomes.

4. **Prescribing and substitution practices:** Healthcare professionals can prescribe generic alternatives, and in some cases, pharmacies may substitute the brand-name drug with a generic equivalent when permitted.

A well-known example within Novartis is *Diovan*, whose active ingredient is valsartan and is used for the treatment of hypertension and heart failure. When Diovan's patent expired in 2012, multiple manufacturers entered the market with generic versions of valsartan, significantly increasing competition and reducing prices.

The Novartis Digital Finance Hub oversees the application of advanced analytics to financial processes. Among its responsibilities is the forecasting of future sales, enabling strategic planning and assessment of the financial impact of events such as generic entries. These forecasts support country organizations in generating monthly and annual sales reports, which are later consolidated and used in company-wide financial accounting.

Context

A drug goes through several stages throughout its lifecycle, from its market introduction to its eventual decline in sales.

When a drug is first launched, its sales volume is typically low, as awareness and adoption are still developing. During the growth phase, sales increase rapidly as the product gains market acceptance, until reaching the maturity phase, where sales tend to stabilize.

After maturity, the entry of new competitors may slow down growth or even trigger a decline in sales volume. Ultimately, once the drug's patent expires—a moment known as the *Loss of Exclusivity* (LoE)—generic versions enter the market. This often leads to a steep and sudden decline in sales volume, a phenomenon known as **generic erosion**. This **generic erosion** is the central topic of this year's Datathon.

From a business perspective, anticipating how a drug will behave after generics enter the market is crucial, as it directly affects revenue forecasts, production planning, and strategic decisions. Accurate erosion predictions enable countries and companies not only to prepare for the post-patent period but also to minimize financial losses and adapt their competitive strategies.

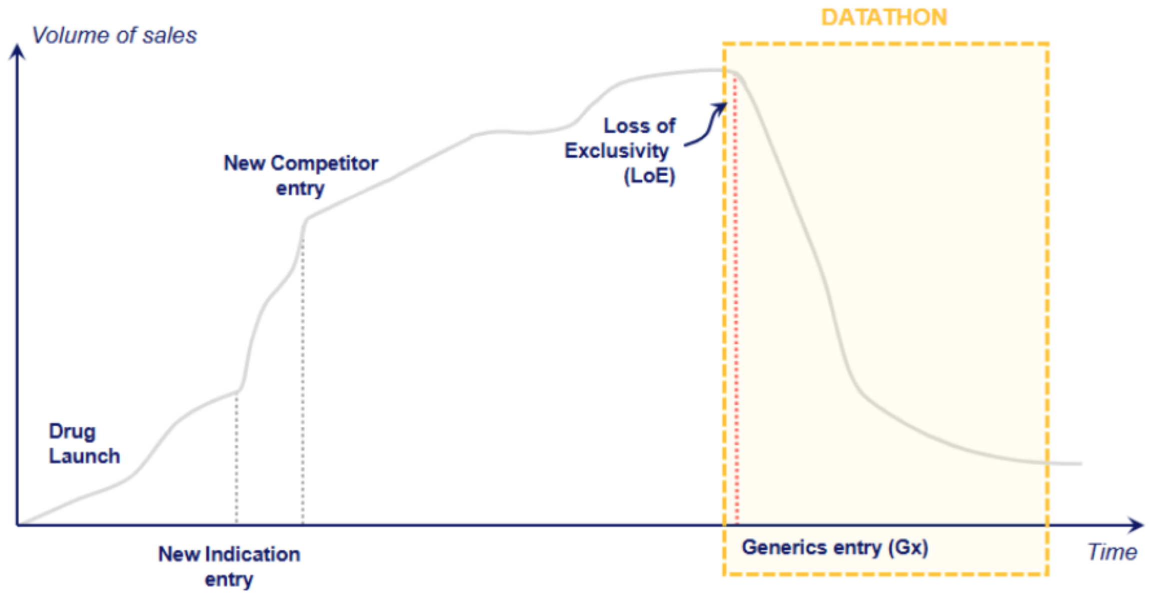


Figure 1.1: Lifecycle of a drug: evolution of sales across different phases.

Generic Erosion

We define the *Mean Generic Erosion* as the average of the normalized volume during the 24 months following the generic entry, as shown in the formula below. Post-entry volumes are normalized using the average monthly volume from the 12 months preceding the entry of generics.

$$\text{Mean Generic Erosion} = \frac{1}{24}\sum_{i=0}^{23} \text{vol}_i^{\text{norm}}, \tag{1.1}$$

$$\text{vol}_i^{\text{norm}} = \frac{\text{Vol}_i}{\text{Avg}_j}, \tag{1.2}$$

$$\text{Avg}_j = \frac{1}{12}\sum_{i=-12}^{-1} Y_{j,i}^{\text{act}}. \tag{1.3}$$

1.2.1 Classification

Drugs may be grouped into three distinct categories depending on their erosion pattern:

- **Low Erosion**: Drugs that experience a minimal impact after the generic entry. In this case, volume remains relatively stable, so the mean normalized erosion stays close to 1.
- **Medium Erosion**: Drugs that show a moderate decline in sales volume after generics enter, with mean erosion levels between 0 and 1.
- **High Erosion**: Drugs that experience a sharp drop in volume, resulting in a mean normalized erosion close to 0.

For the purposes of this Datathon, we classify drugs into two erosion buckets:

- **Bucket 1**: High erosion drugs, with mean erosion between 0 and 0.25. These are the primary focus of the Datathon.
- **Bucket 2**: Drugs with mean erosion greater than 0.25.

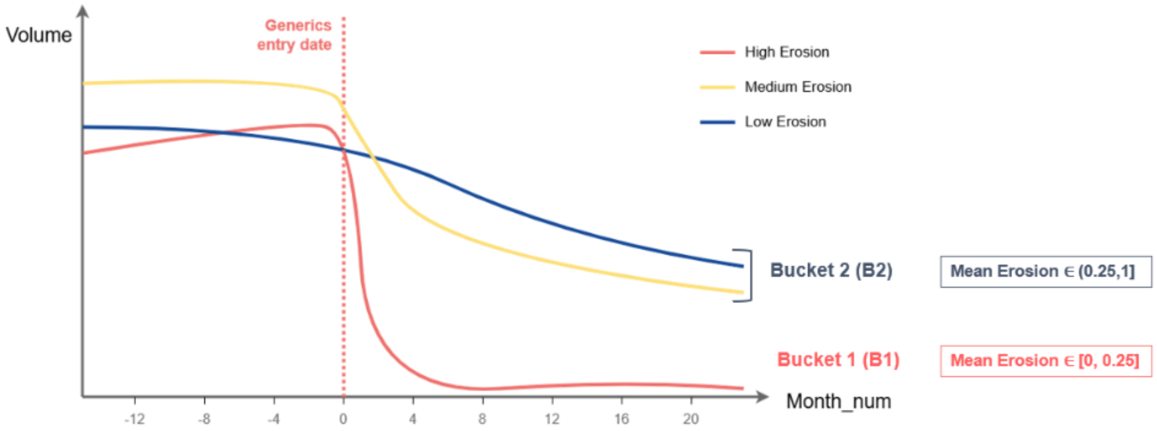


Figure 1.2: Representation of generic erosion classification.

The Challenge

The goal of this Datathon is to forecast the volume erosion that occurs after the entry of generic competitors into the market. Participants are asked to model and predict how the volume of a drug evolves over a 24-month period following the generic entry date.

Forecasts must be produced under two different business scenarios:

- **Scenario 1:** Forecast conducted immediately after the generic entry date, with no observed data post-entry. Participants must forecast monthly volumes from Month 0 to Month 23.
- **Scenario 2:** Forecast conducted six months after the generic entry date. Participants will have access to six months of post-entry actual data and must forecast from Month 6 to Month 23.

Business Challenge

Even though this Datathon challenge has a strong technical component, there is also a **business dimension** that drives the question we're asking participants to solve. So, it's not only about *how* you solve the problem, but also *why* you approach it that way.

Therefore, all teams presenting in front of the jury will be asked to deliver a deep exploratory analysis of their data preprocessing, with a special **focus on high-erosion cases** — that is, drugs whose sales experience a sharp drop after the generic entry.

Finally, we strongly encourage candidates to use visualization tools to make their findings clear, interpretable, and business-oriented.

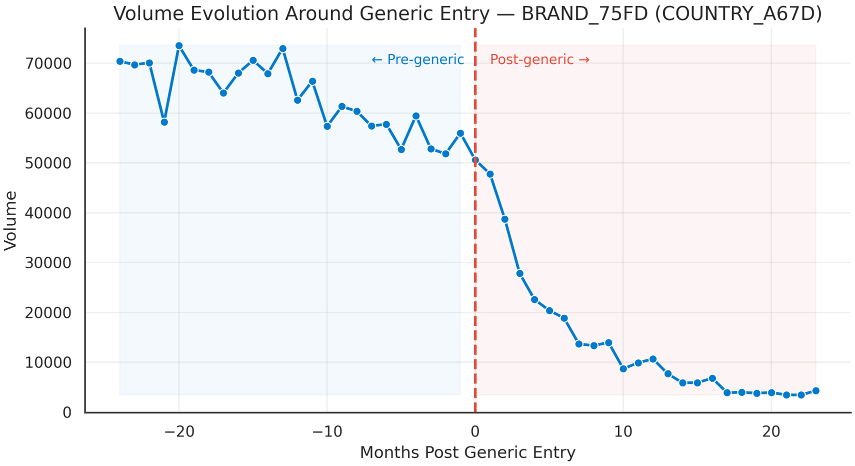


Figure 2.1: Example of volume evolution pre and post generic entry

Evaluation Criteria

The selection of the Datathon winners will follow a two-phase evaluation process.

Phase 1: Model Evaluation

All participating teams must submit volume predictions for the full test dataset, which includes both Scenario 1 and Scenario 2 cases. Phase 1 consists of two sequential steps:

1. **Phase 1-a – Scenario 1 Accuracy**
   All teams will be evaluated based on their forecasting accuracy for Scenario 1. Teams will be ranked according to their prediction errors, and the top 10 teams will advance to Step 2.

2. **Phase 1-b – Scenario 2 Accuracy**
   Only the top 10 teams from Step 1 will be evaluated on their forecasting accuracy for Scenario 2. Based on this second accuracy ranking, the top 5 teams will advance to Phase 2.

Phase 2: Jury Evaluation

The final five teams will present their methodologies, modeling decisions, insights, and conclusions to a jury composed of both technical and business experts. After reviewing all presentations, the jury will select the top three winning teams.

The Data

Participants are provided with datasets containing historical monthly volumes for pharmaceutical products that have experienced generic entry. The **target variable** is monthly volume for 2,293 country-brand combinations that have undergone a generic entry event. The data is organized into train and a test set, as described below.

- **Training Set:** 1,953 observations
  Each observation includes up to 24 months of volume before generic entry and up to 24 months after entry, allowing participants to analyze historical patterns and post-entry dynamics.
- **Test Set:** 340 observations
  The test set is divided into two forecasting scenarios:
  - **Scenario 1 (228 observations):** Forecast from Month 0 to Month 23, immediately following the generic entry.
  - **Scenario 2 (112 observations):** Forecast from Month 6 to Month 23, where the first 6 months of post-entry volume are provided.

Understanding the structure and purpose of these datasets is essential for designing robust forecasting models and tailoring approaches to the two distinct scenarios.

Description of the DataFrames

The data is provided in three separate DataFrames: a volume dataset, a generics dataset, and a dataset containing descriptive information about each medicine.

3.1.1 Volume Dataset (`df_volume.csv`)

This dataset contains sales volumes before and after the entry of generic competitors. Each record includes:

- `country`: market of reference,
- `brand_name`: product of interest,
- `month`: calendar month of the observation,
- `months_postgx`: number of months relative to the month of generic entry (0 = entry month; negative = prior months; positive = subsequent months),
- `volume`: number of units sold.

The variable `volume` is used as the target variable in the analysis.

3.1.2 Generics Dataset (`df_generics.csv`)

This dataset provides information on the presence and evolution of generics for each country–brand combination. It includes:

- `country` and `brand_name`: identifiers,
- `months_postgx`: defined as above,
- `n_gxs`: number of generics available at each point in time.

The number of generics may vary over time as products enter or exit the market.

3.1.3 Medicine Information Dataset (`df_medicine_info.csv`)

This dataset provides contextual characteristics for each drug. It includes:

- `country` and `brand_name`,
- `therapeutic_area`: therapeutic category of the medicine,
- `hospital_rate`: proportion of units distributed through hospitals,
- `main_package`: predominant commercial format (e.g., pills, vials),
- `biological`: boolean indicating whether the product is biological,
- `small_molecule`: boolean indicating whether the product is a low–molecular-weight synthetic compound.

This information supports the characterization of products and the interpretation of observed sales dynamics.

Additional Information

- All historical sales volume is provided at the monthly level, starting from either the brand launch or the first available data point.
- You may use all data provided for model training (e.g., test data from Scenario 2 may be used when training the model for Scenario 1).
- Bucket information is not included in the current dataset but can be derived using Equation 1.1.
- You are free to use any approach or model for your solution; however, explainability and simplicity of the results will be valued.
- Volume may be reported in different units (milligrams, packs, pills, etc.) depending on the country–brand combination.
- Categorical variables can be assumed to remain constant over time.
- Some columns may contain missing values. It is up to you to decide whether to keep them as they are, infer missing values, or apply another preprocessing method.

Metric: Prediction Error

Understanding the evaluation metric is essential for interpreting model performance in this challenge. Forecast accuracy for both scenarios is assessed using a modified version of the Prediction Error (PE), specifically designed to capture short- and long-term deviations between predicted and actual volumes after generic entry. The metric combines monthly errors and accumulated errors across different post-entry periods, normalized for comparability across country–brand combinations.

Metric Phase 1-a

4.1.1 Description

In the first phase (Scenario 1), participants must provide predictions without access to any actual post-entry volume data. To capture early and mid-term erosion patterns, the prediction error is computed from four components with the following weights:

- Absolute monthly error for all 24 months (20%)
- Absolute accumulated error for Months 0–5 (50%)
- Absolute accumulated error for Months 6–11 (20%)
- Absolute accumulated error for Months 12–23 (10%)

All four components are normalized by the average monthly volume of the last 12 months before generic entry (denoted Avg_j), ensuring comparability across brands.

4.1.2 Formula

$$
\begin{aligned}
PE_j = 0.2 \left(\frac{\sum_{i=0}^{23} \left| Y_{j,i}^{act} - Y_{j,i}^{pred} \right|}{24 \cdot Avg_j} \right) + 0.5 \left(\frac{\left| \sum_{i=0}^{5} Y_{j,i}^{act} - \sum_{i=0}^{5} Y_{j,i}^{pred} \right|}{6 \cdot Avg_j} \right) \\
+ 0.2 \left(\frac{\left| \sum_{i=6}^{11} Y_{j,i}^{act} - \sum_{i=6}^{11} Y_{j,i}^{pred} \right|}{6 \cdot Avg_j} \right) + 0.1 \left(\frac{\left| \sum_{i=12}^{23} Y_{j,i}^{act} - \sum_{i=12}^{23} Y_{j,i}^{pred} \right|}{12 \cdot Avg_j} \right)
\end{aligned}
\tag{4.1}
$$

Metric Phase 1-b

4.2.1 Description

In the second phase (Scenario 2), participants receive the first six months of actual post-entry data. Predictions are required from Month 6 onward. Because this phase benefits from partial information, the metric uses three components with the following weights:

- Absolute monthly error for Months 6–23 (20%)
- Absolute accumulated error for Months 6–11 (50%)
- Absolute accumulated error for Months 12–23 (30%)

All components are normalized by the average monthly volume of the last 12 months before generic entry (Avg_j).

4.2.2 Formula

$$\begin{aligned} PE_j = 0.2 \left(\frac{\sum_{i=6}^{23} \left| Y_{j,i}^{act} - Y_{j,i}^{pred} \right|}{18 \cdot Avg_j} \right) + 0.5 \left(\frac{\left| \sum_{i=6}^{11} Y_{j,i}^{act} - \sum_{i=6}^{11} Y_{j,i}^{pred} \right|}{6 \cdot Avg_j} \right) \\ + 0.3 \left(\frac{\left| \sum_{i=12}^{23} Y_{j,i}^{act} - \sum_{i=12}^{23} Y_{j,i}^{pred} \right|}{12 \cdot Avg_j} \right) \end{aligned} \tag{4.2}$$

Final Metric and Interpretation

After computing PE_j for each country–brand combination, scenario-level performance is obtained by aggregating errors across two predefined buckets. Bucket 1 (high erosion) is given twice the weight of Bucket 2:

$$PE = \frac{2}{n_{B_1}} \sum_{j=1}^{n_{B_1}} PE_{j,B_1} + \frac{1}{n_{B_2}} \sum_{j=1}^{n_{B_2}} PE_{j,B_2}. \tag{4.3}$$

Each country–brand pair appears in only one scenario, leading to a distinct (PE_j, equation 4.1 and 4.2) for every observation. These are then aggregated using the weighting shown above to produce a single performance score per scenario.

Ideally, PE_j values lie between 0 and 1, where 0 indicates perfect accuracy. Values greater than 1 are possible and indicate poor prediction quality. At the scenario level, if all $PE_j = 0$, the final score is 0, and if all $PE_j = 1$, the final score equals 3. Because individual errors may exceed 1, scenario-level values can also exceed 3.

Metric Rationale

The Prediction Error metric provides a comprehensive assessment of model performance by incorporating three key dimensions:

- **Generic erosion:** Captured through the bucket-level weighting (Equation. 4.3), where high-erosion cases have double importance.
- **Months since generic entry:** Reflected through differential weighting across time periods in Equations 4.1 and 4.2, emphasizing early post-entry months.
- **Seasonality effects:** Included via the monthly-error terms in both formulas, ensuring that short-term fluctuations contribute appropriately to the final score.

Participants should aim to minimize the final Prediction Error by generating forecasts that accurately capture early erosion dynamics, stabilize over time, and reflect seasonal variation.

The Platform

In this section, we provide details about the communication and submission platforms used for the Datathon, as explained by the Eurecat team.

Communication Platform: Microsoft Teams

Microsoft Teams will serve as the primary platform for all communication. Two main channels will be available: *Mentoring* and *Novartis Datathon*.

The **Mentoring** channel allows private communication between teams and mentors, especially during mentoring meetings. These meetings will last 10 minutes, and teams will be able to book a time slot through a URL shared on Microsoft Teams. Mentors will be responsible for starting these meetings, and teams must join at the agreed time by clicking the *Meet* button.

The **Novartis Datathon** channel is designated for general communication. In the *General* sub-channel, only mentors can post, providing general information. In the *Files* tab, you will find the folder *DATA FOR PARTICIPANTS*, which includes essential folders such as:

- **SUBMISSION**: contains all necessary files for the competition, including data files, metric files for cross-validation, and instructions for uploading results.
- **KICK-OFF-CHALLENGE**: includes additional introductory materials.

A template for preparing the final presentation slides will also be available in the *Files* tab. Organizers will be reachable at the times specified in the Agenda.

Submission Platform

5.2.1 Access

To access the submission platform, participants must log in using the team username and password provided. The link to the platform can be found in the document *submission_instructions*.

5.2.2 Submission Process

Upon entering the platform for the first time, teams must change their password for security purposes. This can be done by navigating to the options on the right, selecting *Profile*, and then *Change password*, as demonstrated in the instructional video.

Once the password is updated, teams may begin uploading submissions. **Important:** submissions must be in `.csv` format. Teams may upload a maximum of **3 submissions**

every 8 hours (see the table below for the time windows), and it's important to be aware that a failed submission does **not** count towards this limit.

Uploading is accomplished by accessing the "Dashboard/Panel" on the left, clicking on "Checkpoint", and using the designated button for submission. If an error message appears, it indicates an issue with the structure of the file.

After the first successful submission, teams will appear in the ranking. The ranking updates after each new submission and will display only the team's best solution.

A **Team Submissions** section allows teams to view and manage their uploaded files. Given the limited number of allowed submissions, it is essential to use each attempt wisely.

8-hour Submission Windows

Please note that all listed times are in **Central European Time** (CET).

Window #	**From**	**To**
Test window after kickoff	Thu 27, 18:00	Thu 27, 20:00
1	Thu 27, 20:00	Fri 28, 04:00
2	Fri 28, 04:00	Fri 28, 12:00
3	Fri 28, 12:00	Fri 28, 20:00
4	Fri 28, 20:00	Sat 29, 04:00
5	Sat 29, 04:00	Sat 29, 12:00
6	Sat 29, 12:00	Sat 29, 20:00
7	Sat 29, 20:00	Sun 30, 04:00
8	Sun 30, 04:00	Sun 30, 10:30 (final submission)

5.2.3 Final Submission

On Sunday at 9:30 a.m., the *Select for final* option will be activated. Teams will have until 10:30 a.m. to choose up to **two submissions** for final evaluation. **The Datathon ends at 10:30 a.m., after which no further changes are allowed.** The metrics will then be calculated on a private portion of the test set.

At 11:30 a.m., the top 10 results will be published, ranked by the first metric. Afterwards, the top 5 solutions, ordered by the second metric, will be announced as finalists.

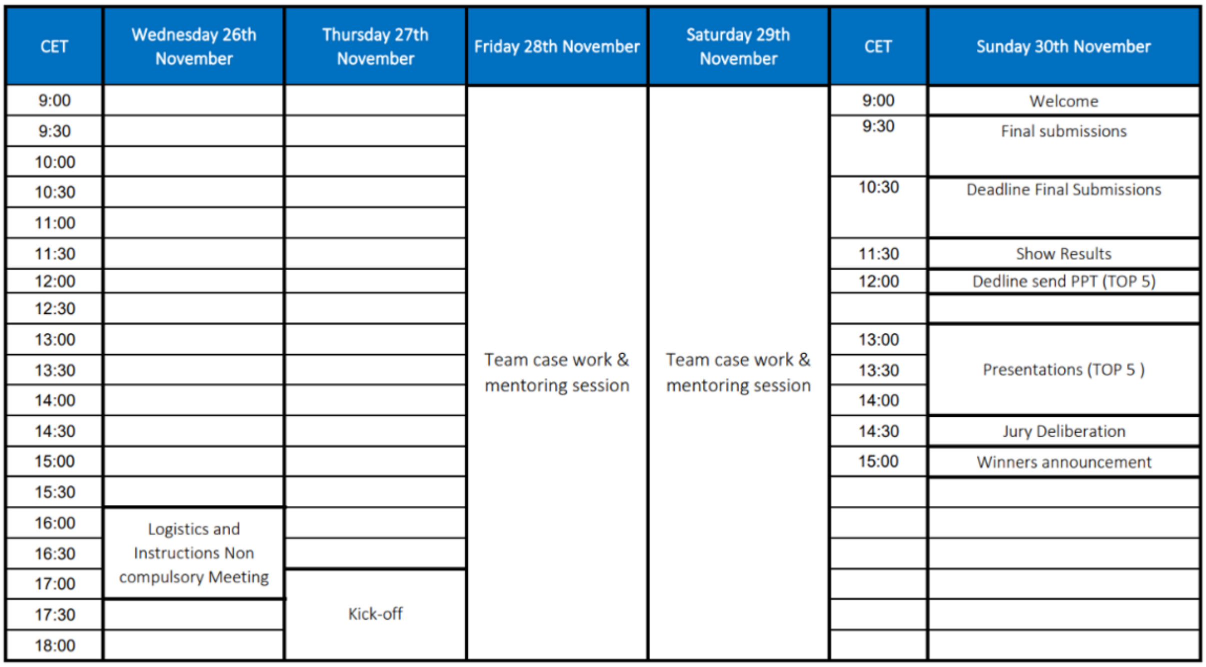
5.2.4 Finalists' Responsibilities

Finalists must prepare a presentation detailing their methodology and results. Presentations must be uploaded by 12:00 p.m. in the private team channel, using the specified naming format. Finalists **must also upload the code** used to generate their final submissions.

Participants are encouraged to prepare their presentation during the Datathon, as time will be limited on the final day.

5.2.5 Jury Presentation

The jury presentations will take place between 13:00 and 14:30. At 15:00, following the jury's deliberation, the winners of the 8th Novartis Datathon will be announced. The final schedule will be provided below.



CET	Wednesday 26th November	Thursday 27th November	Friday 28th November	Saturday 29th November	CET	Sunday 30th November
9:00			Team case work & mentoring session	Team case work & mentoring session	9:00	Welcome
9:30					9:30	Final submissions
10:00						
10:30					10:30	Deadline Final Submissions
11:00						
11:30					11:30	Show Results
12:00					12:00	Dedline send PPT (TOP 5)
12:30						
13:00					13:00	Presentations (TOP 5)
13:30					13:30	
14:00					14:00	
14:30					14:30	Jury Deliberation
15:00					15:00	Winners announcement
15:30						
16:00	Logistics and Instructions Non compulsory Meeting					
16:30						
17:00		Kick-off				
17:30						
18:00						

Figure 5.1: Final Novartis Datathon Schedule