Breast Cancer classification:

CureAllCancers Inc. is a small, cash-constrained biotech start-up founded by Pak (CEO, MD), Basil (CFO, CPA), and Anthony (Chief Medical Officer, MD with limited biostatistics training).

The company is developing a low-cost, point-of-care screening device intended for use in regional clinics and mobile screening units. The device produces a set of numeric measurements derived from digital microscopy images of breast tissue samples.

Key constraints and complications:

- CureAllCancers has no historical internal data.
- Early testing devices were deployed across multiple clinics with inconsistent calibration standards.
- Data collection was rushed to meet investor milestones.
- Measurement protocols evolved during early trials.
- Some clinics manually transcribed values from legacy systems.
- Labels (cancer vs non-cancer) are based on follow-up biopsies, which are sometimes delayed, missing, or disputed.

Due to time and budget constraints, CureAllCancers does not yet have production-quality data. Instead, they have assembled an early proof-of-concept dataset derived from multiple pilot deployments and external reference sources. This dataset is known to contain:

- Measurement noise
- Inconsistent calibration across clinics
- Missing values
- Redundant or poorly documented features
- Possible label uncertainty

At this stage, the dataset is considered frozen.

Management has explicitly decided not to invest further resources in data cleaning or recollection for this phase of the project. Your team's task is not to fix the data, but to work with it as-is and assess what can (and cannot) be reliably inferred.

Objectives:

Your team has been hired to:

1. **Build a classification model** that predicts whether a patient has breast cancer.
2. **Evaluate model performance** under realistic business constraints.

3. **Clearly communicate trade-offs and risks** associated with using this dataset for decision-making.

The output of the model will be a **binary screening decision**:

- **Positive** → patient is flagged for follow-up testing
- **Negative** → patient is not flagged

Business Constraints

CureAllCancers has provided the following guidance:

- False negatives are more costly than false positives
  Missing a cancer case is unacceptable from both medical and reputational perspectives.
- False positives are not free
  Excessive follow-ups increase operational costs and strain clinic capacity.
- The model will not be used for diagnosis, only for screening and triage.

Scope Constraints (Important)

To reflect real-world product timelines:

- You must use the dataset exactly as provided
- No data cleaning, imputation, feature removal, or relabeling is allowed
- You may:
  - Choose models.
  - Choose evaluation metrics.
  - Tune thresholds.
  - Exclude observations *only if explicitly justified as a modeling decision* (e.g., convergence failure, model constraints).

You are encouraged to reach out to Basil for key questions for clarification of the project. Just note that Basil and Pak do not have a data science background and so may only be able to provide more insights on the domain side of the project.

Deliverables for the project:

1. A presentation deck that provides:

a. A description of the data with some preliminary analysis.
b. Data treatment (if any) approaches.
c. Assumptions that were made (mentioned or not mentioned by the client).
d. Model inputs and outputs.
e. Performance metrics and results.

2. Code files used to create files.

Other than technical details (including the replicability of your code), your team will be evaluated on the following criteria

1. The appearance and flow of the slides.
2. The delivery of the presentation of each member → This includes:
   a. Catering to the right audience.
   b. Rhythm of the communication.
   c. Confidence of communicating the results.

A tip to ensure the above items are met well is to carve time to prepare the deck and rehearse several times.