## **Clalit Innovation – Data Science Position: Home Assignment**

Welcome to the home assignment for the Data Science position at Clalit Innovation!

In this assignment, you are provided with a CSV file containing various features (independent variables) and associated health conditions. Your task is to analyze this data and develop predictive models. Please address the following:

### **Core Tasks**

1. **Exploratory Analysis:**Derive a few interesting insights from the data. Highlight trends, anomalies, or potential challenges.
2. **Model Development:**Build at least two different predictive models (using different algorithms) to classify health outcomes (e.g., healthy vs. specific disease categories). Decide on an appropriate metric and try to optimize it to get the best score possible.
3. **Model Explainability:**Use SHAP or a similar tool to explain the predictions of one of your models
4. **Data Improvement Suggestions:**Propose additional data or features that would improve model performance and reliability.

### **Bonus Tasks (Optional)**

* **Containerization:**Package the project in a Docker container for reproducibility.
* **Neural Network Implementation:**Implement one of your predictive models using a neural network architecture.

### **Guidelines**

* We recommend spending no more than **two hours** on this assignment.
* We understand the final model’s metrics may not be perfect — we’re more interested in your thought process, structure, and approach.
* You are welcome to use any tools or resources, including OpenAI or code generation assistants. Please note, the code will be assessed as if you wrote it 100% yourself.
* Please write clean, modular, and maintainable code.
* While the dataset currently contains ~10,000 records, please design your solution as if it contained **1 billion records**.

**Shipping**

Feel free to ship the project in any of the following ways:

1. Email the project to [bakugan@gmail.com](mailto:bakugan@gmail.com)
2. Push it to a GitHub repo and invite/send the link to the email above.

Please note, **all** the data is **completely fake** and is not based on anything real.

P.S. - We would love to hear if you found the explanation intuitive enough, and if not, which part?