Clalit Innovation – NLP Position: Home Assignment

Welcome to the home assignment for the NLP position at Clalit Innovation!

In this assignment, you are provided with a JSON file containing various doctor appointment summaries.

Core Tasks

- Create clusters of diseases according to whatever criteria you see fit. The output should be groups of diseases that are alike.

- Rank doctors by patient outcomes. Detect outlier doctors (good or bad). The output should be a ranking of all doctors by doctor_id.

- Build a prediction model that takes whatever features you choose (from the given data) and predicts whether the future outcome will be better or worse.

Bonus Tasks (Optional)

- **Containerization:**
  Package the project in a Docker container for reproducibility.

Guidelines

- We recommend spending no more than **two hours** on this assignment. Feel free to work on the 3 tasks in whatever order you like. We understand you might not make it to all 3 of them.

- 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 other 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 is quite small, 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
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?