Development of a Custom GenAI Solution for Conducting Health Data Analysis

Background

In the contemporary era of advanced healthcare, understanding the relationship between lifestyle choices, genetics and health outcomes is crucial for personalized medicine. With the rise of data-driven health analytics, it is possible to understand various health conditions by analysing a combination of demographic, genetic, and lifestyle factors.

The dataset provided here offers a rich array of variables on individuals' health metrics and lifestyle choices. However, it's difficult to quickly and accurately interpret and generate context-aware insights out of it. And, for the purpose, there is a need to have a robust and reliable knowledge retrieval mechanism in place.

Note: *The datasets provided here (refer to the next page) are hypothetical.*

Objectives

- The challenge is to develop a GenAI solution that can **effectively retrieve and generate** relevant information from the dataset, providing valuable descriptive insights and recommendations based on the user's queries.
- The model should be capable of **understanding complex queries and integrating information from multiple data sources** to generate accurate and coherent responses. Avoid consolidating the multiple datasets and retrieving information from the same. Instead, make the GenAI solution good enough to join the data on the fly temporarily and retrieve the information from the temporarily consolidated data.
- The focus is on building a system that can integrate diverse data points, generate contextually relevant responses, and **refine its performance through iterative enhancements**.

Deliverables

1. **Data audit report** comprising your understanding of data, based on data preprocessing and preliminary exploratory analysis. **(Optional)**
2. **End-to-end pipeline development**: **(Mandatory)**
    a. **Data extraction** and **preprocessing**, data **integration** and **feature engineering** (if any)
    b. **Integration of a freely available GenAI model** to generate natural language responses/recommendations based on the structured data analysis. Avoid feeding in both the datasets into the LLM as unstructured inputs; instead, develop a **GenAI solution that creates SQL/Python query as an interim output** in the backend, that should **fetch the required subset of dataset for the generation of desired responses**/recommendations on the same in natural language
    c. Model **fine-tuning/instruction-tuning**
    d. **Response generation**, **evaluation** (definition and implementation of metrics & frameworks for evaluating the quality of generated responses) and **refinement**
    e. A **simple web-based interface** where users can input their data and receive contextually relevant responses. Feel free to use free frameworks like *Flask*, *Streamlit or Gradio* to build the interface.
3. **Presentation** explaining your approach, challenges faced, and how your solution addresses the problem. **(Mandatory)**
4. **Provision of software code** used for data preprocessing, model fine-tuning, and evaluation. Provide comprehensive documentation for your pipeline, including setup instructions, code comments, and explanations of design decisions. **(Optional)**

Note:

- **Critically consider ethical factors**, such as data privacy and the implications of providing health recommendations. Some public Generative AI services may leverage user data for future training sets, potentially exposing proprietary data.
- **Your final GenAI solution will be validated** on some queries (to be given to you during the interview).

Datasets and Data Dictionaries

Dataset #1 (attached herewith): Health Dataset 1 (N=2000)

Health Dataset 1 (N=2000).csv

Variable	Position	Variable Label	Value Labels	Measurement Level
Patient_Number	1	Patient Number	Not Applicable	Nominal
Blood_Pressure_Abnormality	2	Blood Pressure Abnormality	0 = Normal 1 = Abnormal	Nominal
Level_of_Hemoglobin	3	Level of Hemoglobin (g/dl)	Not Applicable	Ratio
Genetic_Pedigree_Coefficient	4	Genetic Pedigree Coefficient*	Not Applicable	Ratio
Age	5	Age	Not Applicable	Ratio
BMI	6	BMI	Not Applicable	Ratio
Sex	7	Sex	0 = Male 1 = Female	Nominal
Pregnancy	8	Pregnancy	0 = No 1 = Yes	Nominal
Smoking	9	Smoking	0 = No 1 = Yes	Nominal
salt_content_in_the_diet	10	Salt content in the diet (mg/per day)	Not Applicable	Ratio
alcohol_consumption_per_day	11	Alcohol consumption per day (ml/day)	Not Applicable	Ratio
Level_of_Stress	12	Level of Stress (Cortisol Secretion)	1 = Low 2 = Normal 3 = High	Ordinal
Chronic_kidney_disease	13	Chronic kidney disease	0 = No 1 = Yes	Nominal
Adrenal_and_thyroid_disorders	14	Adrenal and thyroid disorders	0 = No 1 = Yes	Nominal

***Genetic Pedigree Coefficient** (GPC) of an individual for a particular disease is a continuum between 0 and 1, where:*
*GPC **closer to 0** indicates very **distant occurrence** of that disease in her/his pedigree, and*
*GPC **closer to 1** indicates very **immediate occurrence** of that disease in her/his pedigree]*

Dataset #2 (attached herewith): Health Dataset 2 (N=20,000)

Health Dataset 2 (N=20000).csv

Variable	Position	Variable Label	Value Labels	Measurement Level
Patient_Number	1	Patient Number	Not Applicable	Nominal
Day_Number	2	Day Number	Not Applicable	Nominal

Physical_activity	3	Physical activity (no. of steps/day) in the last 10 days	Not Applicable	Ratio