Data Audit Report

Patient data is given in 2 datasets

1. Health data – Contains demographic details like age, sex; habits like smoking, alcohol consumption etc
2. Activity data – Step count for each day

There is missing data and some times it may require special attention. For example, if gender is male then pregnancy column is always not applicable but for females missing value could mean not pregnant currently or status unknown. Similarly, pregnancy is more likely only in a certain age bracket.

We may have to build data quality checks before feeding the raw data into database.

There is also metadata given about tables which is helpful. It gives a more detailed description of column names plus it has encoding key given for fields.

How to handle the data-

For RAG applications, text data is embedded and stored in a vector DB. The given data is tabular in format with no significant long text in it. We can create a Data layer and abstract the underlying data stored in xlsx, csv, sqlite db etc.

Also, the given data has encodings (e.g. in the sex column, 0 means male & 1 means female). We can store the data after converting encodings back to their labels. Or we can store data as given and specify data transformations that transform this data after retrieval and before feeding to LLM.

Project Development

IDE – VS code due to availability of extensions such as connecting to github, being able to work with WSL

OS – Ubuntu running within WSL and terminal connected through VS code which allows easy access to linux commands and less surprizes when moving from development to server deployment

Repo – Code hosted on Github for online backup and version control -> https://github.com/YashKushwaha/chat_ui

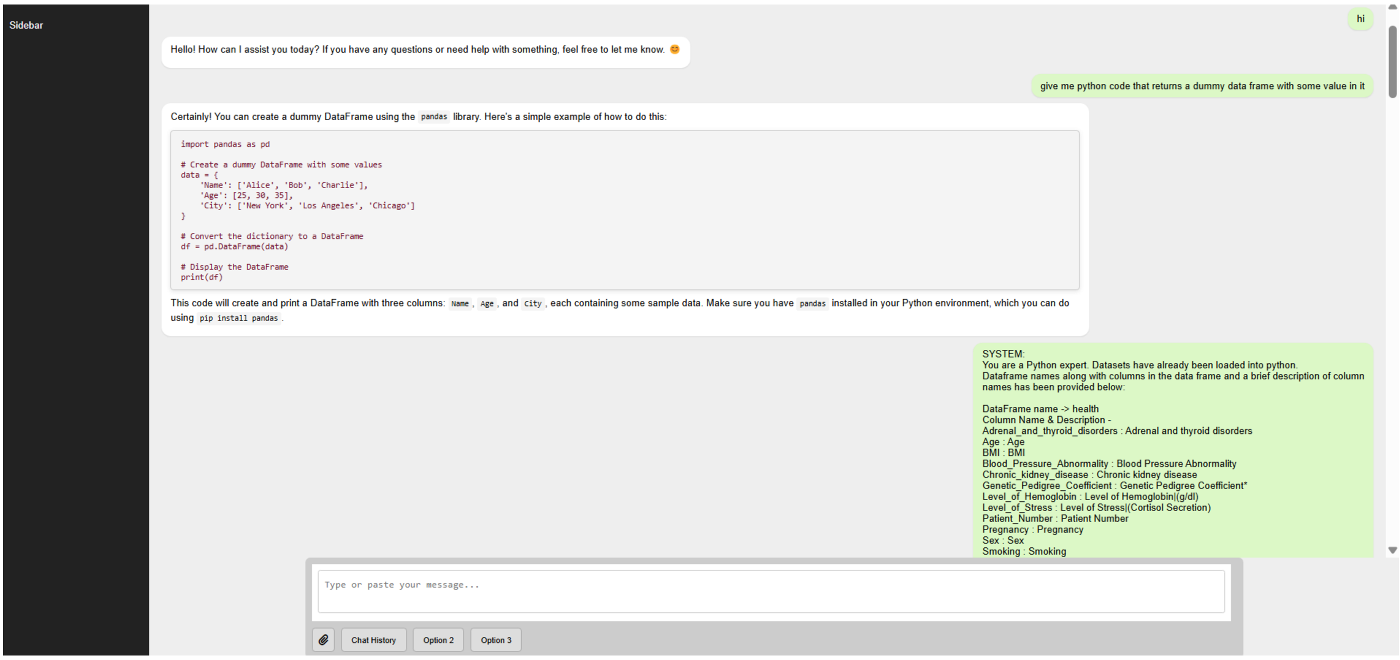
LLM Model – Currently using Phi-4 17B params 4 bit mixed quantization model being served through ollama. Advantage –

- Data remains offline and secure
- Also gives a sense of reasoning capabilities required for the tasks. If current model not capable then we can go for Phi-4 model with more parameters or lesser quantization (e.g 8 bit or 16 bit). We can try other open source models (llmama, mistral etc) or go for commercial LLMs (e.g. ChatGPT)
- Backend – FastAPI used to develop routes. We can have endpoints for simple LLM calls, RAG, Agents and test routes as well
- Front end – Designed using HTML, CSS & simple JS. Web frame works like Node.js not used to keep the interface simple and light weight. Key features -
    - Messaging app like design to distinguish user input and bot/LLM response
    - UI can render markdown as well leading to better formatting plus separate box for codes
    - Sidebar for keeping track of conversation (to be implemented in future)

Architecture Used

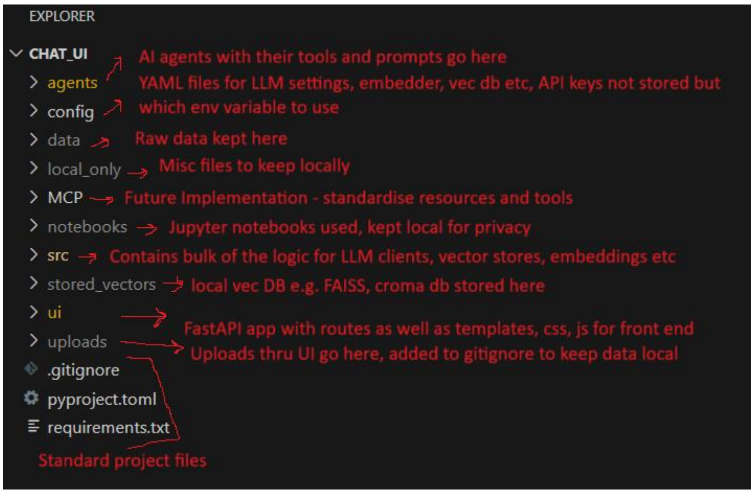
I have used router agent architecture where a master agent receives the user query & then interprets it and classifies the query. The query is then delegated to specialized sub agent best suited for the query. Currently 2 sub agents are implemented –

1) Agent 1 – Takes as input a patient number and then retrieves patient data from the database and analyses the data and creates a report highlighting the finding
2) Agent 2 – Converts user queries into python code which allows for joining multiple tables. This python code is then executed and then analysed by the LLM.



Project Structure

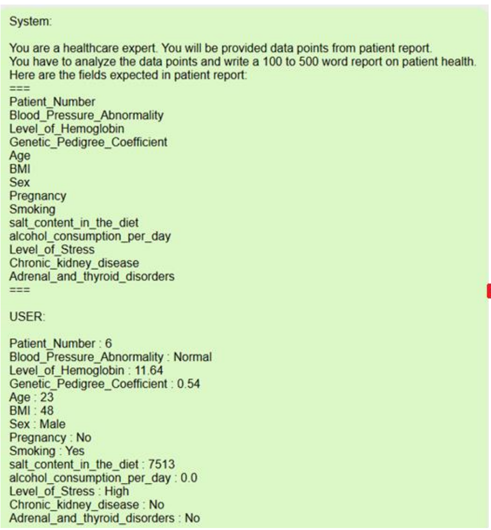
Designed to keep code modular and easy to extend



Evaluation

Manual evaluation to check quality of responses

Check 1 – Can the model understand the data provided? LLM response is good

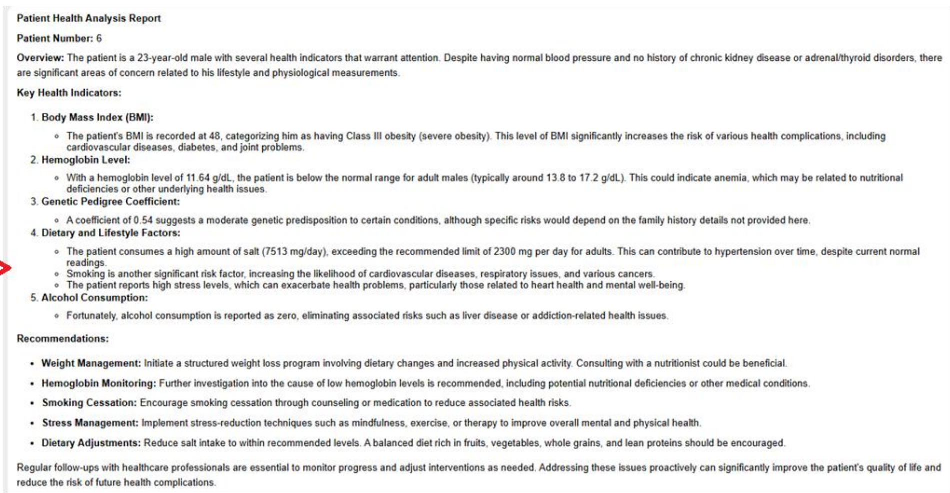


System:

You are a healthcare expert. You will be provided data points from patient report.
You have to analyze the data points and write a 100 to 500 word report on patient health.
Here are the fields expected in patient report:
===
Patient_Number
Blood_Pressure_Abnormality
Level_of_Hemoglobin
Genetic_Pedigree_Coefficient
Age
BMI
Sex
Pregnancy
Smoking
salt_content_in_the_diet
alcohol_consumption_per_day
Level_of_Stress
Chronic_kidney_disease
Adrenal_and_thyroid_disorders
===

USER:

Patient_Number : 6
Blood_Pressure_Abnormality : Normal
Level_of_Hemoglobin : 11.64
Genetic_Pedigree_Coefficient : 0.54
Age : 23
BMI : 48
Sex : Male
Pregnancy : No
Smoking : Yes
salt_content_in_the_diet : 7513
alcohol_consumption_per_day : 0.0
Level_of_Stress : High
Chronic_kidney_disease : No
Adrenal_and_thyroid_disorders : No



Patient Health Analysis Report

Patient Number: 6

Overview: The patient is a 23-year-old male with several health indicators that warrant attention. Despite having normal blood pressure and no history of chronic kidney disease or adrenal/thyroid disorders, there are significant areas of concern related to his lifestyle and physiological measurements.

Key Health Indicators:

1. **Body Mass Index (BMI):**
   - The patient's BMI is recorded at 48, categorizing him as having Class III obesity (severe obesity). This level of BMI significantly increases the risk of various health complications, including cardiovascular diseases, diabetes, and joint problems.
2. **Hemoglobin Level:**
   - With a hemoglobin level of 11.64 g/dL, the patient is below the normal range for adult males (typically around 13.8 to 17.2 g/dL). This could indicate anemia, which may be related to nutritional deficiencies or other underlying health issues.
3. **Genetic Pedigree Coefficient:**
   - A coefficient of 0.54 suggests a moderate genetic predisposition to certain conditions, although specific risks would depend on the family history details not provided here.
4. **Dietary and Lifestyle Factors:**
   - The patient consumes a high amount of salt (7513 mg/day), exceeding the recommended limit of 2300 mg per day for adults. This can contribute to hypertension over time, despite current normal readings.
   - Smoking is another significant risk factor, increasing the likelihood of cardiovascular diseases, respiratory issues, and various cancers.
   - The patient reports high stress levels, which can exacerbate health problems, particularly those related to heart health and mental well-being.
5. **Alcohol Consumption:**
   - Fortunately, alcohol consumption is reported as zero, eliminating associated risks such as liver disease or addiction-related health issues.

Recommendations:

- **Weight Management:** Initiate a structured weight loss program involving dietary changes and increased physical activity. Consulting with a nutritionist could be beneficial.
- **Hemoglobin Monitoring:** Further investigation into the cause of low hemoglobin levels is recommended, including potential nutritional deficiencies or other medical conditions.
- **Smoking Cessation:** Encourage smoking cessation through counseling or medication to reduce associated health risks.
- **Stress Management:** Implement stress-reduction techniques such as mindfulness, exercise, or therapy to improve overall mental and physical health.
- **Dietary Adjustments:** Reduce salt intake to within recommended levels. A balanced diet rich in fruits, vegetables, whole grains, and lean proteins should be encouraged.

Regular follow-ups with healthcare professionals are essential to monitor progress and adjust interventions as needed. Addressing these issues proactively can significantly improve the patient's quality of life and reduce the risk of future health complications.

Check 2 – Can the LLM generate python code to join tables and perform data analysis based on user query? Yes

SYSTEM:
You are a Python expert. You have been given access to following tables:

DataFrame -> health
Adrenal_and_thyroid_disorders : Adrenal and thyroid disorders
Age : Age
BMI : BMI
Blood_Pressure_Abnormality : Blood Pressure Abnormality
Chronic_kidney_disease : Chronic kidney disease
Genetic_Pedigree_Coefficient : Genetic Pedigree Coefficient*
Level_of_Hemoglobin : Level of Hemoglobin|(g/dl)
Level_of_Stress : Level of Stress|(Cortisol Secretion)
Patient_Number : Patient Number
Pregnancy : Pregnancy
Sex : Sex
Smoking : Smoking
alcohol_consumption_per_day : Alcohol consumption per day (ml/day)
salt_content_in_the_diet : Salt content in the diet|(mg/per day)

DataFrame -> activity
Day_Number : Day Number
Patient_Number : Patient Number
Physical_activity : Physical activity (no. of steps/day) in the last 10 days

Your job is to convert user query into python statements that can be executed.

USER:
How do I get the list of male patients who smoke but also walk a lot every day ?

To retrieve a list of male patients who smoke and take a lot of steps each day, we need to perform the following operations:

1. Filter the `health` DataFrame for male patients who smoke.
2. Determine what constitutes "a lot of steps" by analyzing the `activity` DataFrame (e.g., you might choose an average or specific threshold).
3. Merge the two filtered DataFrames on `Patient_Number` to find common entries that satisfy both conditions.

Below is a Python code snippet using pandas to achieve this:

```
import pandas as pd

# Assuming 'health' and 'activity' are pre-loaded pandas DataFrame objects

# Step 1: Filter male patients who smoke from the health DataFrame
male_smokers = health[(health['Sex'] == 'Male') & (health['Smoking'] == True)]

# If smoking is stored in a different format, you might need to adjust this condition.

# Step 2: Determine what constitutes "a lot of steps"
# This example assumes that "a lot" means more than the 75th percentile of daily steps
threshold_steps = activity['Physical_activity'].quantile(0.75)

# Filter patients in the activity DataFrame who walk a lot every day
active_patients = activity[activity['Physical_activity'] > threshold_steps]

# Step 3: Merge filtered DataFrames on 'Patient_Number' to get male smokers who are also active
result = pd.merge(male_smokers, active_patients, on='Patient_Number')

# Optionally, drop duplicates if a patient appears more than once in the activity data
result = result.drop_duplicates(subset=['Patient_Number'])

# Get the list of unique Patient Numbers or full rows depending on your need
male_smoking_active_patients_list = result['Patient_Number'].unique()

print(male_smoking_active_patients_list)
```

Note:

- Adjust the threshold for "a lot" based on your specific criteria.
- Ensure that `health` and `activity` DataFrames are properly loaded before executing this code. If necessary, you can load them using pandas read functions like `pd.read_csv()`.
- The condition `(health['Smoking'] == True)` assumes 'Smoking' is a boolean column; adjust according to the actual data format (e.g., if it's categorical or coded differently).

Check 3 – Test the entire pipeline where Agent takes the input, retrieves data internally and then returns the final output.

can you tell me about the health of patient number 20 ?

Based on the provided patient report, here's an analysis of the key data points:

1. **Blood Pressure Abnormality**: The patient has abnormal blood pressure readings. This could be related to several factors including obesity (BMI of 38), high stress levels, and chronic kidney disease.
2. **Level of Hemoglobin**: At 16.55 g/dL, this level is slightly elevated for an 18-year-old male (normal range: approximately 13.8-17.2 g/dL). This could be indicative of dehydration or polycythemia, but further tests would be needed to confirm.
3. **Genetic Pedigree Coefficient**: A coefficient of 0.91 suggests a strong genetic predisposition to certain conditions, possibly related to blood pressure abnormalities, kidney disease, or other inherited disorders.
4. **BMI**: A BMI of 38 classifies the patient as obese (Class II obesity). This is a significant risk factor for hypertension, cardiovascular diseases, and can exacerbate chronic kidney disease.
5. **Chronic Kidney Disease (CKD)**: The presence of CKD is critical. It can contribute to hypertension and may be related to adrenal and thyroid disorders, which are also present in this patient.
6. **Adrenal and Thyroid Disorders**: These conditions can affect blood pressure regulation and metabolic processes. They could be contributing to the abnormal blood pressure readings and possibly affecting hemoglobin levels.
7. **Dietary Salt Content**: The high salt intake (1807 mg per day) is concerning, as it can exacerbate hypertension, especially in individuals with CKD or adrenal disorders.
8. **Alcohol Consumption**: At 0.0 g/day, alcohol consumption is not a contributing factor to the patient's current health issues.
9. **Level of Stress**: High stress levels can contribute to elevated blood pressure and may worsen chronic conditions like kidney disease.
10. **Other Factors**: The patient does not smoke or have other lifestyle factors that might contribute to these conditions, which narrows down potential external causes for some of the observed abnormalities.

Recommendations:

- **Medical Management**: Regular monitoring of blood pressure, renal function tests, and hormone levels (adrenal and thyroid) is crucial. Medication may be necessary to manage hypertension, CKD, and any endocrine disorders.
- **Lifestyle Modifications**: A significant reduction in dietary salt intake is recommended. Implementing a weight loss plan with dietary changes and increased physical activity could help reduce BMI and improve overall health.
- **Stress Management**: Techniques such as mindfulness, counseling, or stress management programs might be beneficial to manage high stress levels.

Check 4 – Check the pipeline for converting user query to Python code and executing it

Do patients with less physical activity have higher BMI ?

The analysis of the provided dataset shows that there is a very weak correlation between BMI (Body Mass Index) and physical activity levels among patients. The correlation coefficient calculated for these two variables is approximately 0.035, which suggests almost no linear relationship between them.

In simpler terms, based on this data, changes in physical activity levels are not strongly associated with changes in BMI. Therefore, the evidence from this dataset does not support the conclusion that patients with less physical activity have a higher BMI.

However, it's important to note a few considerations:

- The correlation coefficient of 0.035 is very close to zero, indicating a negligible relationship.
- Correlation does not imply causation; even if there were a stronger correlation, other factors might influence both BMI and physical activity levels.
- Additional analysis could be beneficial to explore non-linear relationships or to consider other variables that might affect BMI, such as diet, age, gender, etc.

If you have further questions or need additional analyses, feel free to ask!

Code generated by the LLM and run in the backend

```python
def run():
    # Step 1: Aggregate physical activity by patient number
    avg_activity_per_patient = activity.groupby('Patient_Number')['Physical_activity'].mean().reset_index()

    # Rename column for clarity after aggregation
    avg_activity_per_patient.rename(columns={'Physical_activity': 'Average_Physical_Activity'}, inplace=True)

    # Step 2: Merge with health data
    merged_data = pd.merge(health, avg_activity_per_patient, on='Patient_Number', how='inner')

    # Step 3: Analyze the relationship between BMI and physical activity
    correlation = merged_data[['BMI', 'Average_Physical_Activity']].corr()

    return correlation
```