Machine Learning Retrieval-Augmented Generation (RAG) System for Personalized Medical Analysis

Yiğit SERT
Sakarya University, Faculty of Computer and Information Sciences
yigit.sert@ogr.sakarya.edu.tr

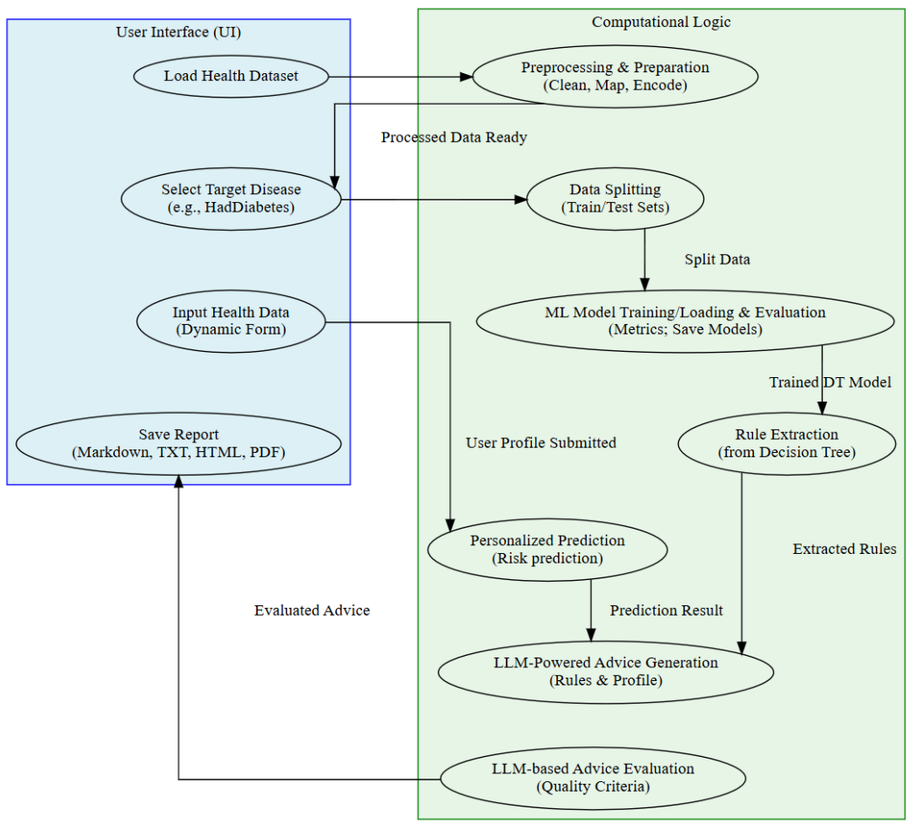
Academic Advisor
Prof. Dr. Devrim AKGÜN

Introduction

This study introduces a system that marries interpretable machine learning with LLMs to deliver personalized health advice. It extracts human-readable if-then rules from a Decision Tree trained on health data, then uses a RAG approach to craft tailored, actionable, and safe recommendations. Implemented with an interactive Gradio interface, the system offers a complete data-to-advice workflow. This innovation addresses key challenges in healthcare: the lack of trust in "black box" ML models, the ineffectiveness of generic advice, and the "last mile" problem of simplifying complex health insights for patients.

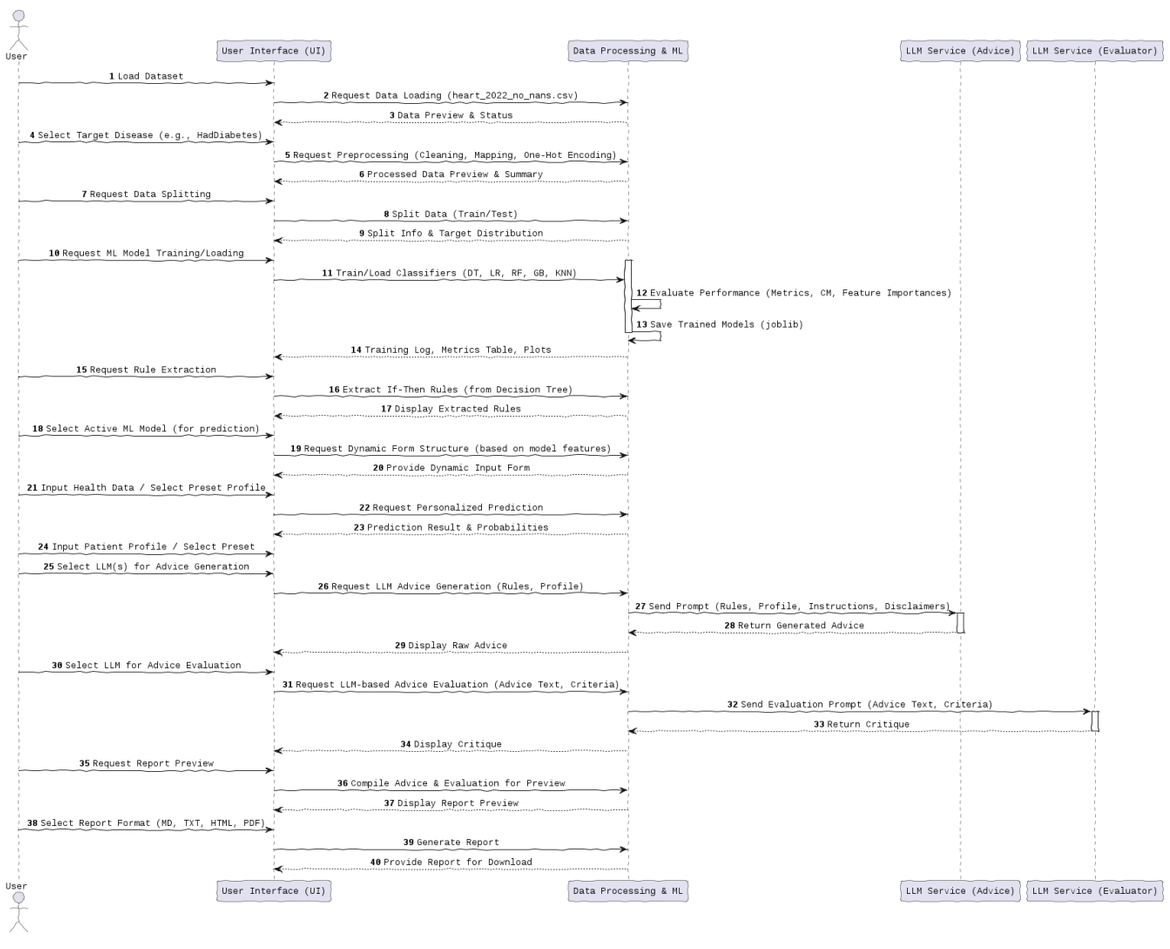
System Architecture

The system is built on a two-part architecture: an interactive User Interface (UI) and a backend Computational Logic engine. The backend manages all data processing, from cleaning and one-hot encoding to training a suite of machine learning models (e.g., Decision Tree) for risk prediction. Critically, it extracts if-then rules from the trained model. These rules, combined with user data, form the context for the Retrieval-Augmented Generation (RAG) process, where Large Language Models (LLMs) are prompted to generate and evaluate personalized advice. The UI, built with Gradio, provides a step-by-step tabbed interface for the user to control this entire process and visualize the results.



The RAG-based Health Advisor Workflow

The user follows a sequential, multi-stage workflow guided by the application's tabbed interface. The workflow begins with setting up and preparing the health dataset, which is automated by the system. Next, the user trains the machine learning classifiers on a selected health target, after which interpretable if-then rules are automatically extracted from the trained Decision Tree model. In the core step, the user provides a patient profile, which is combined with the extracted rules to generate personalized health advice from one or more LLMs. Finally, the generated advice and any accompanying evaluation are compiled into a comprehensive report that can be saved in multiple formats (PDF, HTML, TXT).



Model Training & Selection

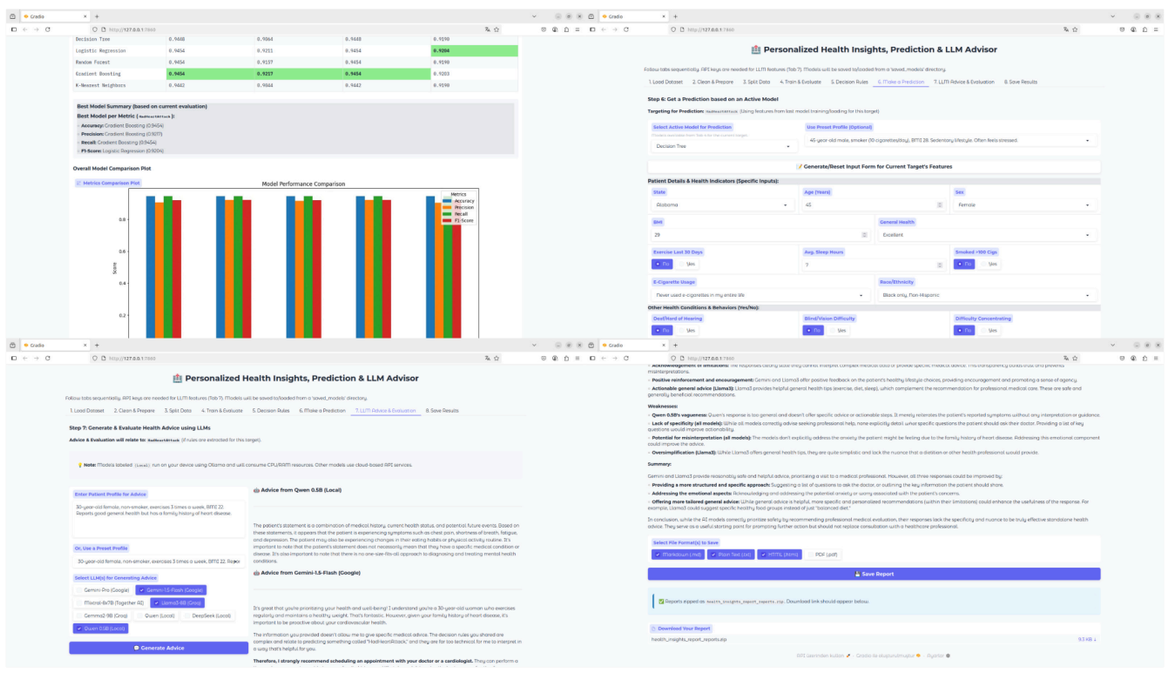
Multiple classifiers were trained to predict health outcomes. The Decision Tree was chosen for its inherent interpretability, which is essential for our rule extraction process.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	**0.8325**	**0.7572**	**0.8325**	0.7584
Logistic Regression	0.8299	0.7423	0.8299	0.7619
Random Forest	0.8097	0.7353	0.8097	0.7630
Gradient Boosting	0.8316	0.7345	0.8316	0.7578
K-Nearest Neighbors	0.8152	0.7377	0.8152	**0.7648**

An insightful metric here is the F1-Score, which balances Precision and Recall and is more robust to class imbalance. The K-Nearest Neighbors (0.7648) and Random Forest (0.7630) models achieve the highest F1-Scores. This indicates they provide a better balance between correctly identifying true positive cases and minimizing false alarms, making them more reliable for some specific tasks.

Interactive User Interface

The system is implemented with a multi-tab web interface using the Gradio library. This design guides the user sequentially through the entire workflow: from data loading and model training to generating and evaluating AI-powered health advice. Key features include dynamic input forms for prediction, live updates of model performance metrics, and a comparative display of advice from multiple LLMs, creating a transparent user experience.



LLM Performance: Speed vs. Quality

Multiple LLMs were evaluated for advice generation. Groq-based models (Llama3, Gemma2) offered near-instant responses, while models like Gemini-1.5-Flash produced more detailed and consistently high-quality advice, as rated by an LLM Judge.

LLM Advice Generator	Average Time (s)	Average Flesch Ease	Average FK Grade
Mixtral-8x7B (Together AI)	6.06	49.5	10.5
Llama3-8B (Groq)	1.25	38.0	13.3
Gemini-1.5-Flash (Google)	4.96	45.0	10.6
Gemma2-9B (Groq)	**0.97**	**53.3**	**9.5**

Note: Averages are calculated across four distinct user scenarios: General Wellness Seeker, Diabetic with Heart Concerns, Young Adult w/ Family History, and Elderly with Arthritis. Time (s) refers to Advice Generation Time. FK Grade is Flesch-Kincaid Grade Level.

Conclusion

A RAG system has been successfully developed to translate intricate ML decision rules into personalized health advice. This system's foundation is a Decision Tree model, which provides a transparent and interpretable basis for the LLM's prompts. An interactive user interface effectively demonstrates a feasible end-to-end workflow, catering to non-expert users. A notable observation during development was the inherent trade-off among LLMs: achieving faster generation speed often necessitates a compromise on the depth and quality of the advice produced.

Future Work

Future improvements include expanding to more datasets, incorporating multi-assessor evaluation, and optimizing the RAG context selection.