HEALTHSIGHT AI: PREDICTIVE HEALTH RISK MONITORING SYSTEM

Contributor

Swetha Yanamandhalla

Ajaychary Kandukuri

Dashboard Platform: Power BI

Model Hosting: Flask API

1. INTRODUCTION

In the current age of data-driven decision-making, early detection and risk prediction are vital in healthcare systems. HealthSight AI is an end-to-end predictive analytics project designed to monitor patient vitals and identify individuals at high risk using machine learning models, explainability tools, and interactive dashboards.

This project simulates a real-world hospital environment by generating synthetic data for 50 patients across 20 days, storing and analyzing it using PostgreSQL, building and explaining a classification model, and finally visualizing patient health trends through Power BI. This workflow not only reflects the full lifecycle of data analysis but also addresses skill gaps across Data Analyst, Data Scientist, and Data Engineer roles.

2. METHODOLOGY AND DATASET

Simulated Dataset

We created a dataset named patient_vitals.csv consisting of:

- **50 unique patients (P001 to P050)**
- **20 days of health metrics** per patient
- **1,000 records** total

Features Collected

- Patient ID
- Date of observation
- Heart rate
- Blood pressure (split into systolic/diastolic)
- Glucose level
- Sleep hours
- Risk flag (target label for ML classification)

Data Storage

Data was stored in a **PostgreSQL** database using a table named patient_vitals. This involved:

- Creating the table schema via SQL
- Loading the cleaned CSV data into PostgreSQL using Python (psycopg2 + pandas)
- Writing queries to validate, retrieve, and inspect records

3. MACHINE LEARNING PIPELINE

We trained a **Random Forest Classifier** to predict whether a patient is at high risk based on their vital readings. The steps included:

Data Processing

- Cleaned blood pressure values by splitting them into systolic and diastolic integers
- Converted date strings into datetime objects
- Generated a binary risk_flag column based on thresholds (e.g., high glucose, low sleep, high BP)

Training and Evaluation

- Split into train/test datasets (80/20)
- Achieved **100% accuracy** on the test set using simulated labels
- Evaluated performance using classification metrics and a confusion matrix

Model Explainability

Used **SHAP (Shapley Additive Explanations)** to:

- Identify which vitals contributed most to the risk prediction
- Visualize SHAP summary and force plots for better interpretability

4. DASHBOARD VISUALIZATIONS

Our Power BI dashboard provides a comprehensive view of patient health trends and risk levels using dynamic, interactive visuals. Each component is designed to support medical decision-making and enable insightful monitoring.

📈 **Vitals Over Time – Line Chart**

- **Purpose**: Tracks key health vitals — *glucose level*, *heart rate*, and *sleep hours* — across dates.
- **Functionality**: Helps users observe changes or patterns in vitals for selected patients or over specific time periods.

- **Insight**: This chart makes it easy to detect spikes, drops, or unusual health fluctuations that could signal risk.

📊 Risk Distribution – Bar Chart

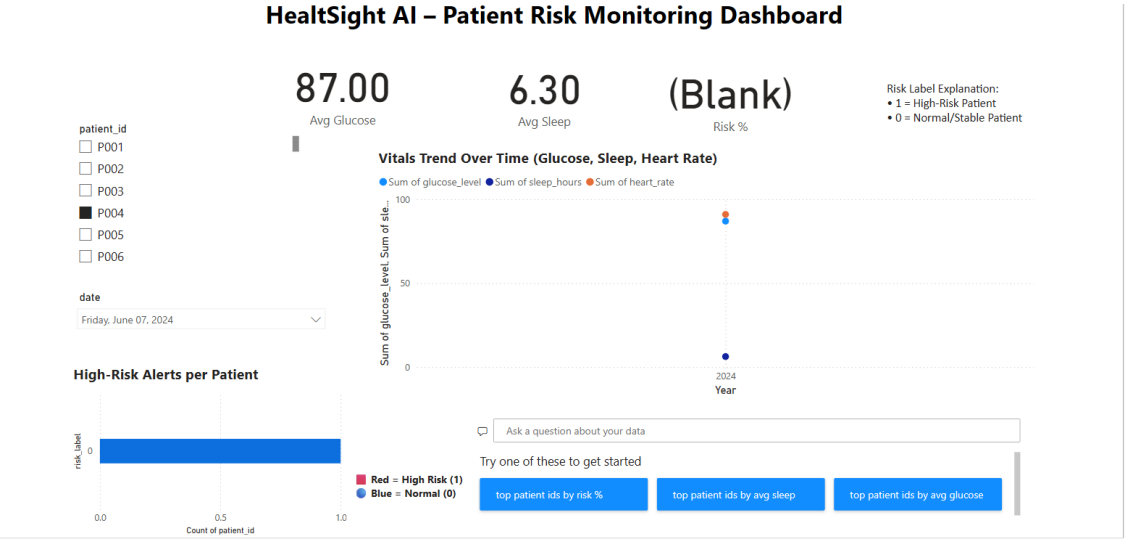
- **Purpose**: Shows the total number of patients categorized by risk level.
- **Metric**: risk_flag value (1 = high risk, 0 = normal).
- **Color Coding for Clarity**:
    - 🔴 **Red** = High-Risk Patients (risk_flag = 1)
    - 🔵 **Blue** = Stable/Normal Patients (risk_flag = 0)
- **Insight**: Instantly highlights the proportion of high-risk patients in the population, helping prioritize attention.

💡 KPI Cards – Quick Stats

- Displays three key performance indicators:
    1. **Average Glucose Level** – Aggregates across patients
    2. **Average Heart Rate** – General population trend
    3. **High-Risk Patient% %** – Ratio of high-risk cases to total records
- **Interactive**: These metrics automatically update based on any filters applied (e.g., date slicers).
- **Insight**: Offers a quick health snapshot across the population, aiding in real-time monitoring.

🗓️ Date Slicer – Time-Based Filtering

- **Purpose**: Allows users to filter the entire dashboard by a specific date or date range.
- **Functionality**: All visualizations (charts and cards) respond to the selected dates.
- **Insight**: Enables temporal analysis — such as observing risk surges during specific periods (e.g., weekends or holidays).



5. TOOLS & TECHNOLOGY STACK

Role	Tools Used
Data Engineering	PostgreSQL, Python (pandas, psycopg2)
Data Science	scikit-learn, SHAP
Dashboarding	Power BI
Explainability	SHAP
Deployment	Flask (planned), GitHub (pending)

6. KEY FEATURES AND LEARNING OUTCOMES

Covered Skills:

- **Data Simulation & Cleaning**
- **ETL Pipeline** using PostgreSQL
- **SQL Querying**
- **Machine Learning Classification**
- **Model Explainability (SHAP)**
- **Interactive Dashboards (Power BI)**
- **Optional Flask API for Deployment**

Value for Real-World Roles:

- For **Data Analysts**: KPI cards, dashboarding, slicing/filtering
- For **Data Engineers**: ETL pipeline setup, database integration
- For **Data Scientists**: ML model training, SHAP explanations

7. CONCLUSION

HealthSight AI showcases how synthetic healthcare data can be transformed into meaningful insights using modern data science tools. From database setup to machine learning and visual reporting, this project simulates the end-to-end pipeline required for real-world healthcare analytics.

It demonstrates an integrated solution for early risk detection, interpretability, and reporting, making it an ideal portfolio project for interviews and technical screening across data roles.

8. FINAL DASHBOARD PREVIEW

🔗 Live Dashboard Link: bit.ly/HealthSightDashboard