Cloud Computing and Big Data Analytics - Final Project

HealthShield – Your Smart Doctor

0. the member list & relative contribution of members

Teamates	Contribution
313657003 Chou Chia Hsuan	Data collection and preprocessing; logistic regression and SVM models; user input interface; final project summary (Sections 1, 3, 6, and 7); presentation slides (Sections 1–3).
M093088 Shih Hsin Yi	Data collection and preprocessing; random forest and XGBoost models; prediction result page; final project summary (Sections 2, 4, 5, and 7); presentation slides (Sections 3–5).

1. Brief Background

Diabetes has become an increasingly serious public health issue in recent years. In Taiwan, diabetes **was ranked among the top five leading causes** of death in 2024. In addition, epidemiological data **from 2015 to 2020 indicate that the incidence of diabetes among younger adults aged 20 to 40 increased by approximately 30% over the five years**. This growing trend highlights the importance of early risk assessment.

2. Motivation

Despite Taiwan's universal health insurance system and relatively high accessibility to medical care, diabetes is often underdiagnosed or diagnosed at a late stage. Many individuals remain unaware of their elevated risk until complications develop. This highlights a critical need for accessible, data-driven tools that can help individuals assess their diabetes risk early and encourage preventive actions before irreversible health damage occurs.

3. Problem Definition

This project aims to **develop a cloud-based machine learning system to predict an individual's risk probability of diabetes using personal information**, including body measurements, medical history, lifestyle habits, blood pressure records, and blood test data.

In addition, interpretability plots are used to illustrate how input features influence an individual's predicted diabetes risk.

The goal is not to replace clinical diagnosis, but to provide a proof-of-concept system that demonstrates how population health data and cloud technologies can be integrated to support early risk awareness and preventive healthcare.

4. Novelty

Compared to RclabRiskPred(NHRI), this project have some novelty:

First, this project shifts diabetes risk prediction from pure classification accuracy to **personalized risk interpretation and clinical usability**. Second, a **SHAP-based explanation module** provides individual-level risk factor analysis, transforming black-box models into interpretable precision medicine tools. Third, the system delivers an **end-to-end, user-facing cloud service** that outputs both risk estimates and personalized explanations. Fourth, a containerized, modular architecture separates frontend, backend, and inference services, aligning with real-world cloud deployment practices.

5. Datasets, Tools, and Algorithms

(1) Datasets

Consider the data accessibility and the completeness of target features, we used NHANES 2007~2018 dataset and chose 1 target and 26 features based on domain knowledge:

Feature Category	Variable Names	Description
Target	diabetes	whether doctor told you have diabetes or not
Basic Information	age	participant's age in years at the time of the survey
	gender	biological sex of the participant, recorded as male or female
Body Measurements	height	standing height measured in centimeters during the physical examination
	weight	body weight measured in kilograms during the physical examination
	bmi	a standardized measure of body adiposity calculated as weight (kg) divided by height squared (m^2)
	waist circumference	waist circumference measured in centimeters, reflecting central (abdominal) adiposity
Blood Test Results	fasting_glucose	plasma glucose concentration measured after an overnight fast, indicating short-term glycemic status
	insulin	fasting serum insulin level, reflecting insulin secretion and insulin resistance
	HbA1c	glycated hemoglobin percentage, representing average blood glucose levels over the past 2–3 months
	total_cholesterol	total concentration of cholesterol in the blood
	HDL	"good" cholesterol associated with protective cardiovascular effects
	LDL	"bad" cholesterol associated with atherosclerosis and cardiovascular disease risk
	triglycerides	blood triglyceride concentration, reflecting lipid metabolism and cardiometabolic risk

Blood Pressure	systolic_1/2/3	repeated measurements of systolic blood pressure (mmHg)
	diastolic_1/2/3:	repeated measurements of diastolic blood pressure (mmHg)
Family History & Lifestyle	ever_smoked	indicator of whether the participant has ever smoked cigarettes
	alcohol_drinks	self-reported alcohol consumption, measured as number of drinks over a specified time period
	vigorous_activity	participation in vigorous-intensity physical activity that substantially increases heart rate or breathing
	moderate_activity	participation in moderate-intensity physical activity that substantially increases heart rate or breathing
	family_diabetes	whether a first-degree family member has been diagnosed with diabetes
	general_health	self-rated overall health status, categorized from excellent(1) to poor(5)
	Sleep_Hours	average number of hours of sleep per night, self-reported by the participant

(2) Data Preprocessing

Before handling missing values of features, we first dropped the missing targets and those samples that were infants, and then randomly split the data into training set and test set with the ratio 80%: 20%.

I. Missing Values

(I) Drop: Missing target, age = 0 (Infant).

(II) Impute with median: waist, fasting_glucose, insulin, HbA1c, total_cholesterol, HDL, triglycerides, LDL, general_health, Sleep_Hours, alcohol_drinks(impute 0 if age<20, o.w. add a new class)

(III) Add a new class: ever_smoked(impute 2(no) if age<20, o.w. add new class), vigorous_activity, moderate_activity, family_diabetes

(IV) Related columns 1: BMI, height, weight－imputed by BMI's formula, or imputed by median.

(V) Related columns 2: systolic_1/2/3, diastolic_1/2/3－take average of the three columns if a column exists value, o.w. imputed by median.

II. Scaling and Encoding

(I) Min-Max normalization (for continuous / multi-level variables): age, bmi, waist_cm, height_cm, weight_kg, fasting_glucose, insulin, HbA1c, total_cholesterol, HDL, triglycerides, LDL, alcohol_drinks, general_health, systolic_avg, diastolic_avg, Sleep_Hours

(II) One-hot encoding (for 2 to 3-class discrete variables): gender, ever_smoked,vigorous_activity, moderate_activity, family_diabetes,

After feature selection and data preprocessing, the remaining number of features = 22, and the remaining number of samples = 56463, where 45170 samples are in the training set and 11293 samples are in the test set.

(3) Tools

I. Python: Used to train models and to build the content of backend/frontend.

II. GCP: Used to develop a cloud-based project.

(4) Algorithm

I. Criteria

We found that in each set, the amount of having diabetes: the amount of not having = 91.66%: 8.34%. To prevent the imbalanced distribution from affecting the judgement of the result, we chose balanced accuracy as our final criterion.

II. Model

Since we aim to predict a binary classification problem, we apply logistic regression, SVM, random forest, XGBoost with cross validation to search for the best parameter.

6. Results

(1) Performance—Best Parameters

Model	Best parameter	Fix Parameters
Logistic Regression	C:128	Penalty: L1, L2, solver=liblinear
SVM	C: 256 Gamma: 0.03125	Kernel: rbf
Random Forest	max_depth: 15, max_features: 'sqrt', max_leaf_nodes: 40, min_samples_leaf: 2, min_samples_split: 5, n_estimators: 300	
XGboost	colsample_bytree: 0.8798, gamma: 0.4645,	

	learning_rate: 0.0553, max_depth: 3, n_estimators: 268, subsample: 0.9801	

(2) Performance－Criteria

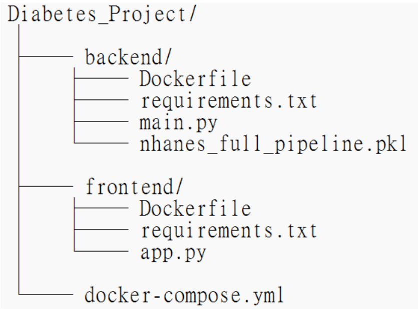
Model	**CV - Training**				**CV - Validation**				Testing			
	BA	Sens	Spec	Accu	BA	Sens	Spec	Accu	BA	Sens	Spec	Accu
LR - L2	0.8894	0.8877	0.8910	0.8907	**0.8881**	0.8854	0.8909	0.8904	**0.8954**	0.9002	0.8906	0.8914
SVM	0.9000	0.9041	0.8959	0.8966	**0.8915**	0.8883	0.8948	0.8942	**0.8945**	0.8938	0.8952	0.8951
RF	0.9082	0.9183	0.8981	0.8998	**0.8958**	0.8949	0.8966	0.8965	**0.8939**	0.8970	0.8907	0.8913
Xgb	0.9140	0.9203	0.9077	0.9087	**0.9015**	0.8970	0.9060	0.9052	**0.9028**	0.9023	0.9033	0.9032

(3) Conclusion

XGBoost performed the best with a balanced accuracy of 0.9028, hence we **choose XGBoost** as our prediction model being used in our cloud platform.

7. Sample Runs & Screenshots

(1) Structure



```
Diabetes_Project/
│
├──── backend/
│      ├──── Dockerfile
│      ├──── requirements.txt
│      ├──── main.py
│      └──── nhanes_full_pipeline.pkl
│
├──── frontend/
│      ├──── Dockerfile
│      ├──── requirements.txt
│      └──── app.py
│
└──── docker-compose.yml
```

(2) Cloud Architecture and Implementation

I. 3 Logical Tiers

Based on the 3-tier architecture theory covered in the course lectures, this project structures the codebase into the following modules, deployed in isolation via Docker containers:

(I) Presentation Logic (**Frontend Container**)

- Implementation Technology: Python Streamlit `frontend/app.py`.
- Functionality: Responsible for **receiving user input health data** (page 1) and **visualizing** the JSON data returned from the backend into **interactive charts** (page 2). This tier contains no prediction logic and is solely responsible for the User Interface presentation.

(II) Business Logic (**Backend Container**)

- Implementation Technology: Python FastAPI `backend/main.py`.
- Functionality: Encapsulates the core prediction algorithms. Upon system startup, it loads the `nhanes_full_pipeline.pkl` model file and **calculates diabetes risk probabilities** using the `predict_proba` method. It also utilizes the SHAP library to **compute feature contributions**.

(III) Data Base Access Logic

- Implementation Form: Serialized model object `backend/nhanes_pipeline_XGBoost.pkl`.
- Functionality: This project uses a **pre-trained XGBoost model** file as a static knowledge base. The backend reads data from this tier using the `joblib` library to perform inferences.

II. **RESTful API** Communication Interface

The frontend and backend containers communicate via the HTTP protocol, adhering to the REST style:

(I) **Interface Definition**: The backend exposes a POST /predict endpoint.

(II) **Data Exchange**: The frontend encapsulates user input into a JSON object for transmission. The backend returns a JSON response containing the prediction probability and a Base64 encoded image.

(III) **Statelessness**: Each API request contains all the information necessary to complete the computation. The server does not retain client state. This characteristic allows the backend containers to automatically scale out via Cloud Run at any time without disrupting the service.

III. Virtualization & Cloud Deployment

(I) Containerization Technology (OS-level Virtualization): Following the containerization concepts, we utilized **Dockerfile to define the execution environment** (based on the python:3.11-slim image). This ensures equivalence between the development and cloud production environments, effectively resolving dependency conflict issues.

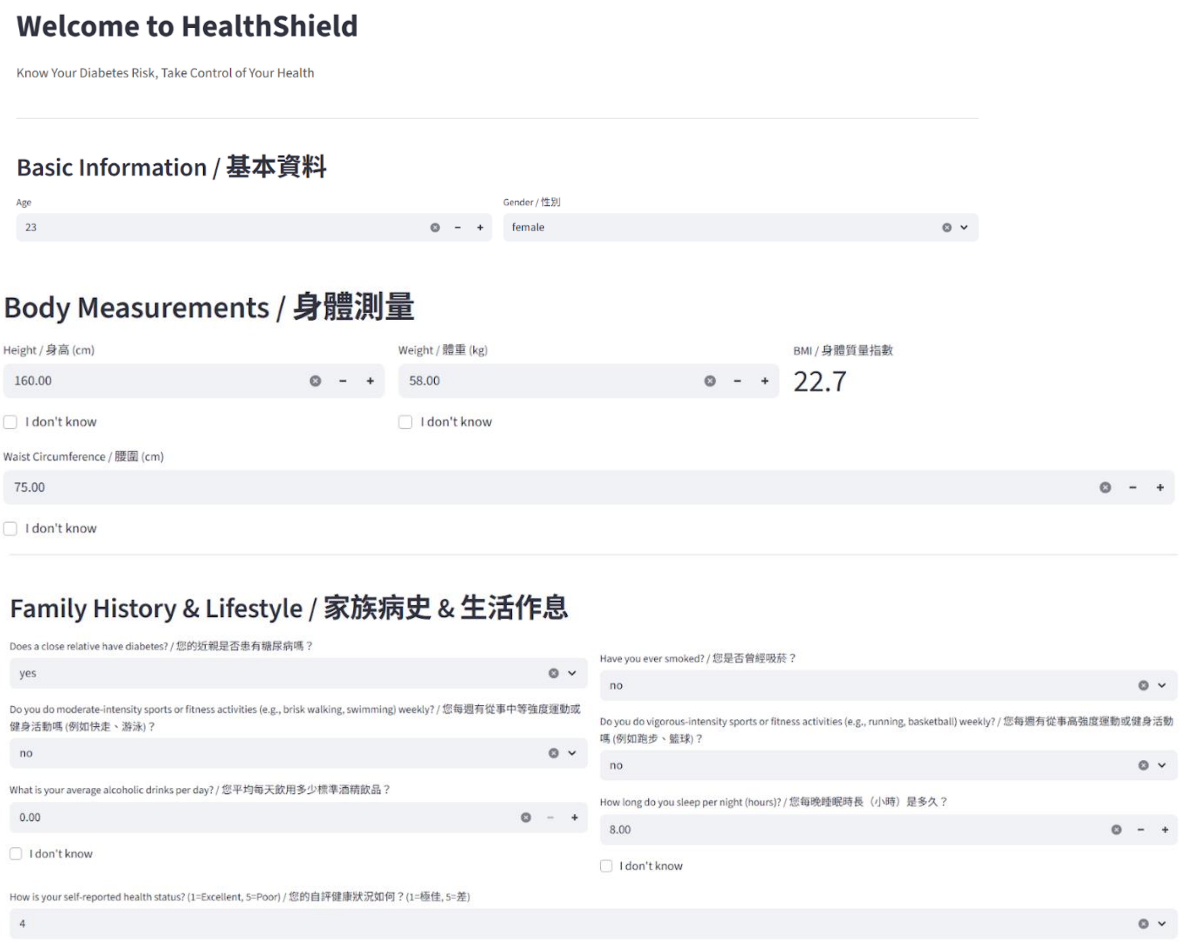
(II) PaaS Platform Application: The project is deployed on **Google Cloud Run**. This is a **PaaS** offering that relieves us from managing the underlying IaaS, allowing us to focus solely on the logic development of main.py and app.py. Cloud Run automatically allocates computing resources based on traffic, realizing the **rapid elasticity** and **pay-per-use** characteristics.

(3) Sample Runs

I. User data input page

The User Data Input Page HealthShield serves as the main interface for users to provide personal health information required for diabetes risk prediction. This page collects various types of input data, including basic personal information, body measurements, family history, lifestyle habits, blood pressure records, and blood test data.

First, users fill in basic information, body measurements, family history, and lifestyle habits, as shown in the corresponding screenshot.



Welcome to HealthShield

Know Your Diabetes Risk, Take Control of Your Health

Basic Information / 基本資料

Age: 23

Gender / 性別: female

Body Measurements / 身體測量

Height / 身高 (cm): 160.00

□ I don't know

Weight / 體重 (kg): 58.00

□ I don't know

BMI / 身體質量指數: 22.7

Waist Circumference / 腰圍 (cm): 75.00

□ I don't know

Family History & Lifestyle / 家族病史 & 生活作息

Does a close relative have diabetes? / 您的近親是否患有糖尿病嗎？: yes

Do you do moderate-intensity sports or fitness activities (e.g., brisk walking, swimming) weekly? / 您每週有從事中等強度運動或健身活動嗎 (例如快走、游泳)？: no

What is your average alcoholic drinks per day? / 您平均每天飲用多少標準酒精飲品？: 0.00

□ I don't know

Have you ever smoked? / 您是否曾經吸菸？: no

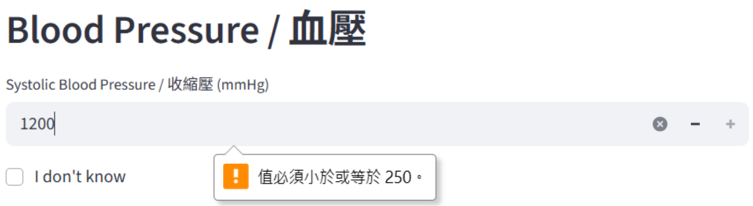
Do you do vigorous-intensity sports or fitness activities (e.g., running, basketball) weekly? / 您每週有從事高強度運動或健身活動嗎 (例如跑步、籃球)？: no

How long do you sleep per night (hours)? / 您每晚睡眠時長（小時）是多久？: 8.00

□ I don't know

How is your self-reported health status? (1=Excellent, 5=Poor) / 您的自評健康狀況如何？(1=極佳, 5=差): 4

Input validation rules and warning messages are implemented to prevent invalid or unreasonable values from being entered. In this screenshot, users have to provide correct blood pressure records before proceeding.

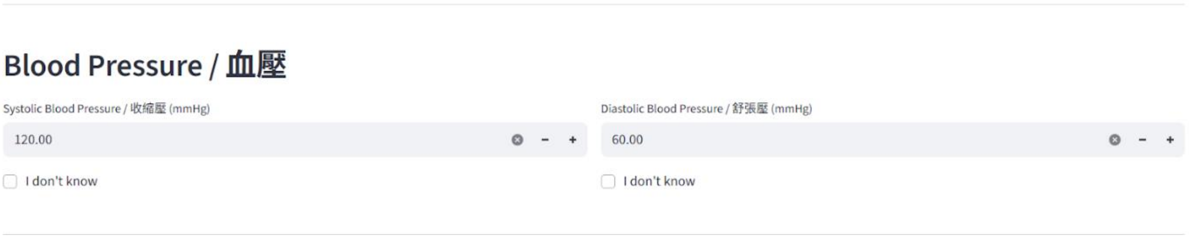


Blood Pressure / 血壓

Systolic Blood Pressure / 收縮壓 (mmHg): 1200

□ I don't know

! 值必須小於或等於 250。



Blood Pressure / 血壓

Systolic Blood Pressure / 收縮壓 (mmHg)

120.00

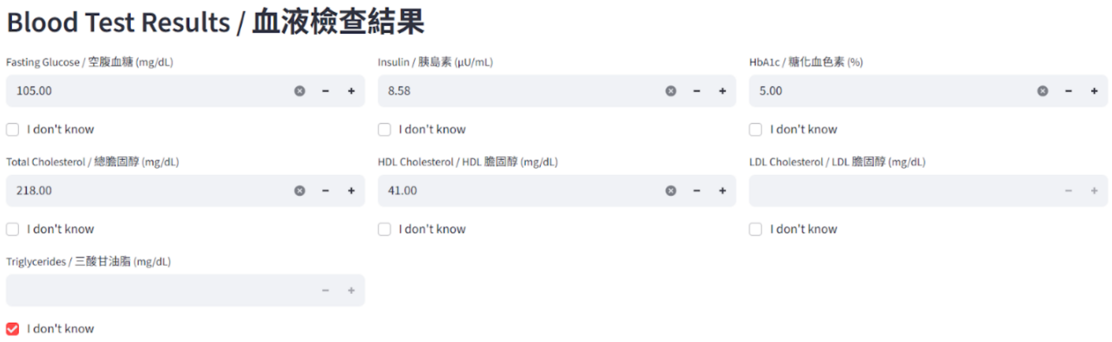
□ I don't know

Diastolic Blood Pressure / 舒張壓 (mmHg)

60.00

□ I don't know

Since some input may not be available to all users, the system provides an “I don’t know” option across multiple data categories, such as blood test results. This design prevents users from guessing unknown values and helps ensure more reliable input data



Blood Test Results / 血液檢查結果

Fasting Glucose / 空腹血糖 (mg/dL)

105.00

□ I don't know

Insulin / 胰島素 (μU/mL)

8.58

□ I don't know

HbA1c / 糖化血色素 (%)

5.00

□ I don't know

Total Cholesterol / 總膽固醇 (mg/dL)

218.00

□ I don't know

HDL Cholesterol / HDL 膽固醇 (mg/dL)

41.00

□ I don't know

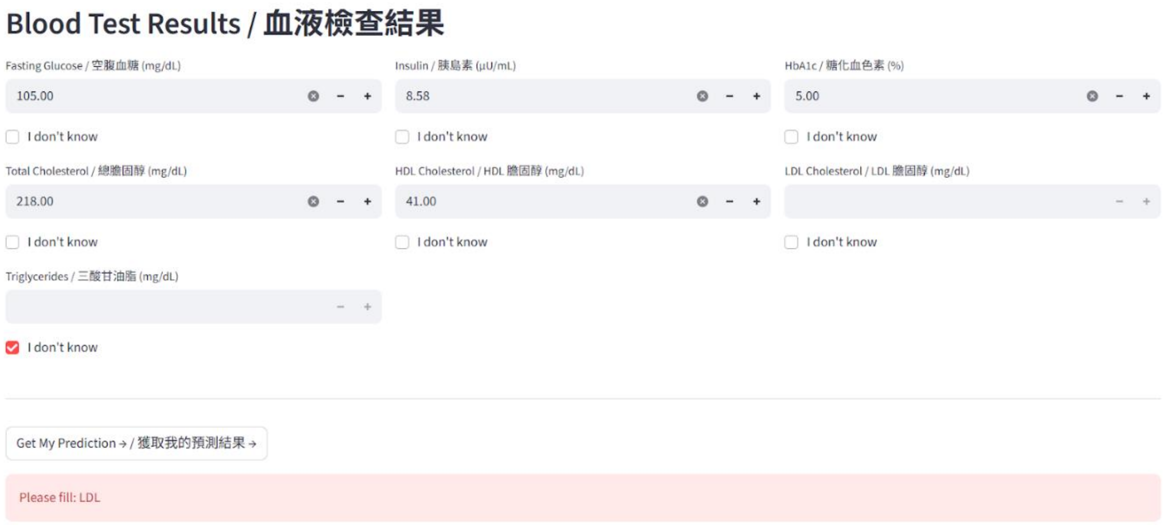
LDL Cholesterol / LDL 膽固醇 (mg/dL)

□ I don't know

Triglycerides / 三酸甘油脂 (mg/dL)

☑ I don't know

In addition, the system performs completeness checks on required fields. If information is missing, warning messages are displayed below the button and cannot proceed to the next page until the necessary information is provided.



Blood Test Results / 血液檢查結果

Fasting Glucose / 空腹血糖 (mg/dL)

105.00

□ I don't know

Insulin / 胰島素 (μU/mL)

8.58

□ I don't know

HbA1c / 糖化血色素 (%)

5.00

□ I don't know

Total Cholesterol / 總膽固醇 (mg/dL)

218.00

□ I don't know

HDL Cholesterol / HDL 膽固醇 (mg/dL)

41.00

□ I don't know

LDL Cholesterol / LDL 膽固醇 (mg/dL)

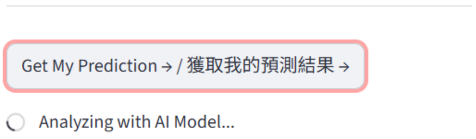
□ I don't know

Triglycerides / 三酸甘油脂 (mg/dL)

☑ I don't know

Please fill: LDL

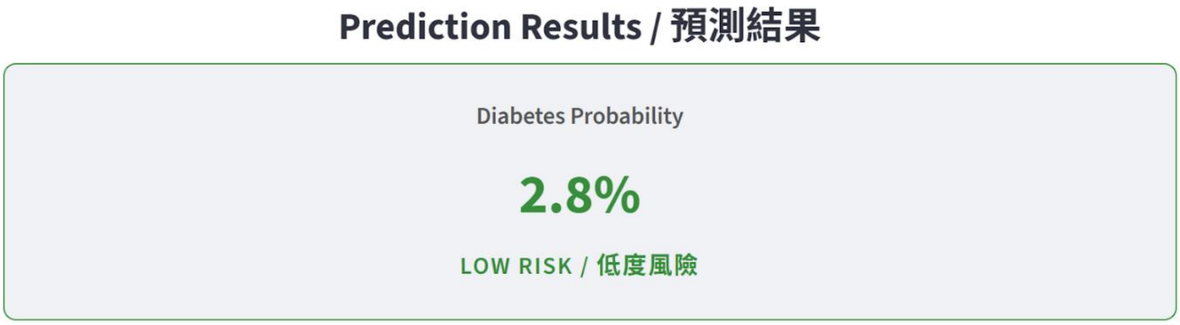
After completing all required inputs, users can submit the form to proceed to the diabetes risk prediction result page.



Analyzing with AI Model...

II. Prediction page

This page presents the diabetes risk prediction generated by the machine learning model. The system displays the predicted diabetes risk probability along with a corresponding risk category (e.g., low, medium, or high risk), allowing users to easily interpret their personalized prediction results.



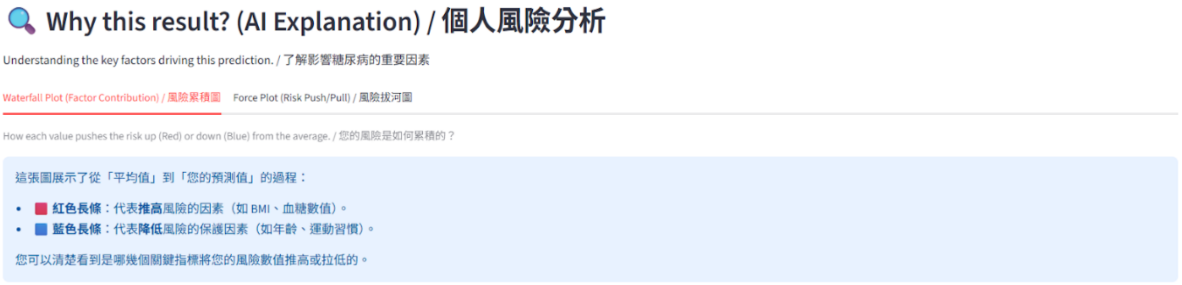
Prediction Results / 預測結果

Diabetes Probability

2.8%

LOW RISK / 低度風險

To provide individual-level interpretability, SHAP-based visualizations are used to explain the model's predictions. The waterfall plot illustrates how each input feature contributes to shifting the prediction from the baseline risk to the final predicted probability for a specific user.



🔍 Why this result? (AI Explanation) / 個人風險分析

Understanding the key factors driving this prediction. / 了解影響糖尿病的重要因素

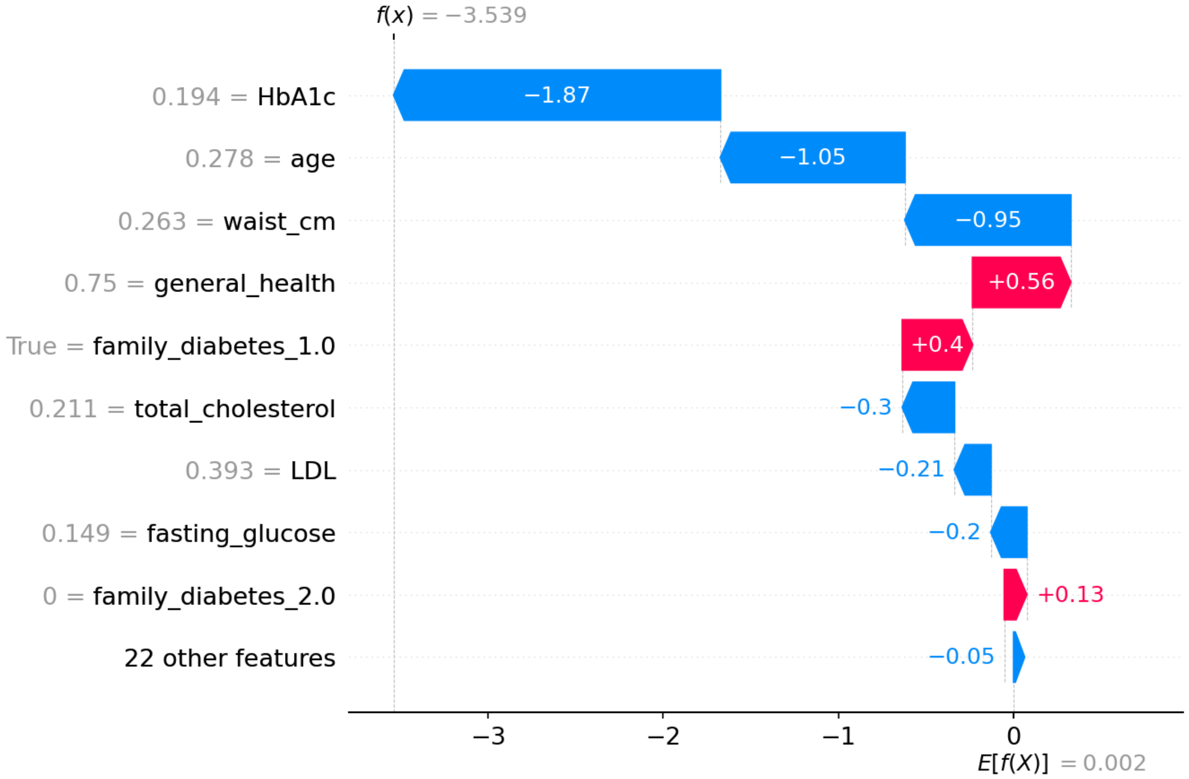
Waterfall Plot (Factor Contribution) / 風險累積圖
Force Plot (Risk Push/Pull) / 風險拔河圖

How each value pushes the risk up (Red) or down (Blue) from the average. / 您的風險是如何累積的？

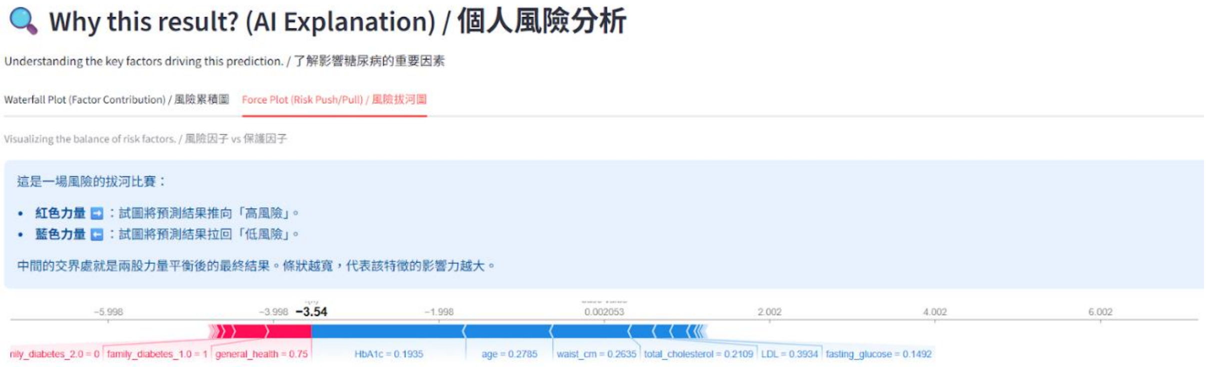
這張圖展示了從「平均值」到「您的預測值」的過程：

- 🟥 **紅色長條**：代表**推高**風險的因素（如 BMI、血糖數值）。
- 🟦 **藍色長條**：代表**降低**風險的保護因素（如年齡、運動習慣）。

您可以清楚看到是哪幾個關鍵指標將您的風險數值推高或拉低的。



In addition, the force plot provides an intuitive summary of the combined effects of all input features, highlighting the overall balance between risk-increasing and risk-reducing factors for the individual prediction.



🔍 Why this result? (AI Explanation) / 個人風險分析

Understanding the key factors driving this prediction. / 了解影響糖尿病的重要因素

Waterfall Plot (Factor Contribution) / 風險累積圖
Force Plot (Risk Push/Pull) / 風險拔河圖

Visualizing the balance of risk factors. / 風險因子 vs 保護因子

這是一場風險的拔河比賽：

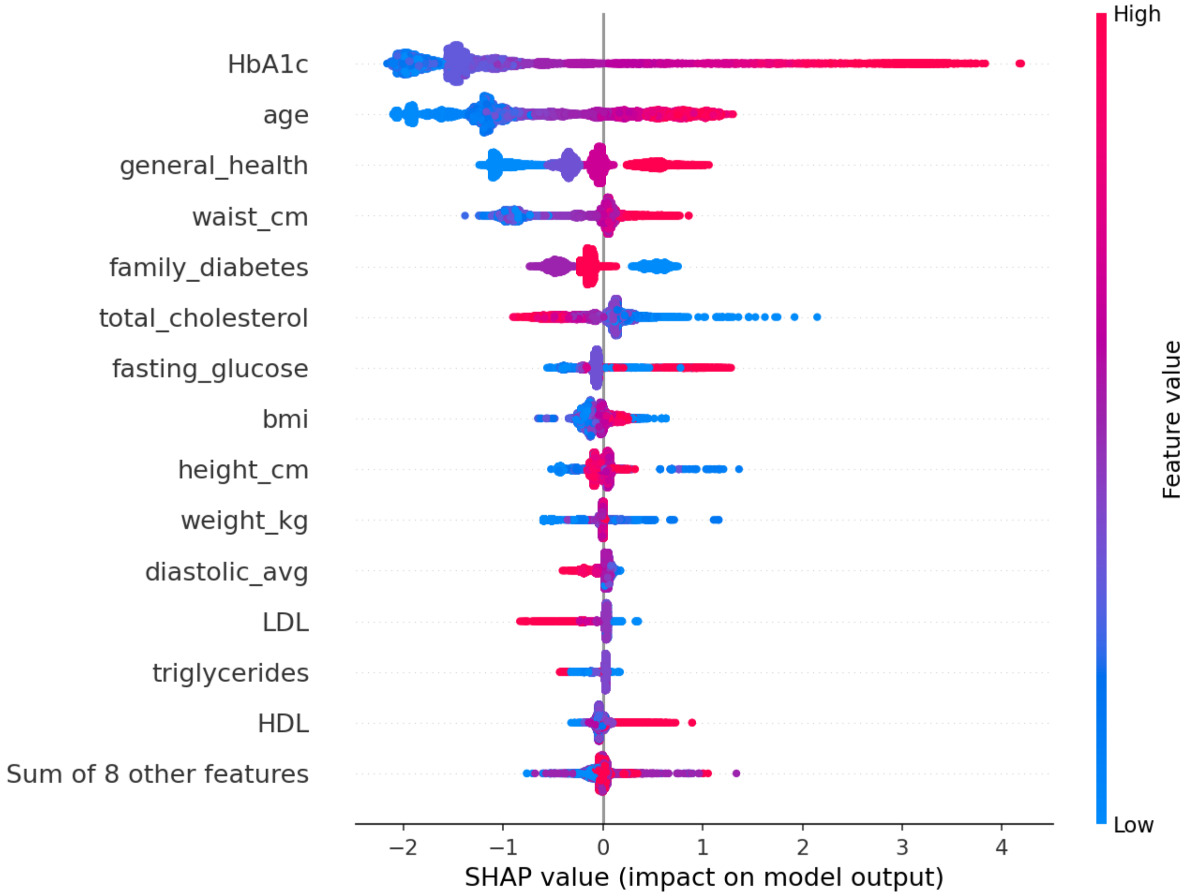
- 紅色力量 ➡：試圖將預測結果推向「高風險」。
- 藍色力量 ⬅：試圖將預測結果拉回「低風險」。

中間的交界處就是兩股力量平衡後的最終結果。條狀越寬，代表該特徵的影響力越大。

To analyze feature importance at a global level, the beeswarm plot visualizes the distribution and impact of input features across the dataset, indicating whether each feature tends to increase or decrease the predicted diabetes risk.

📊 Global Explanation / 模型整體解釋

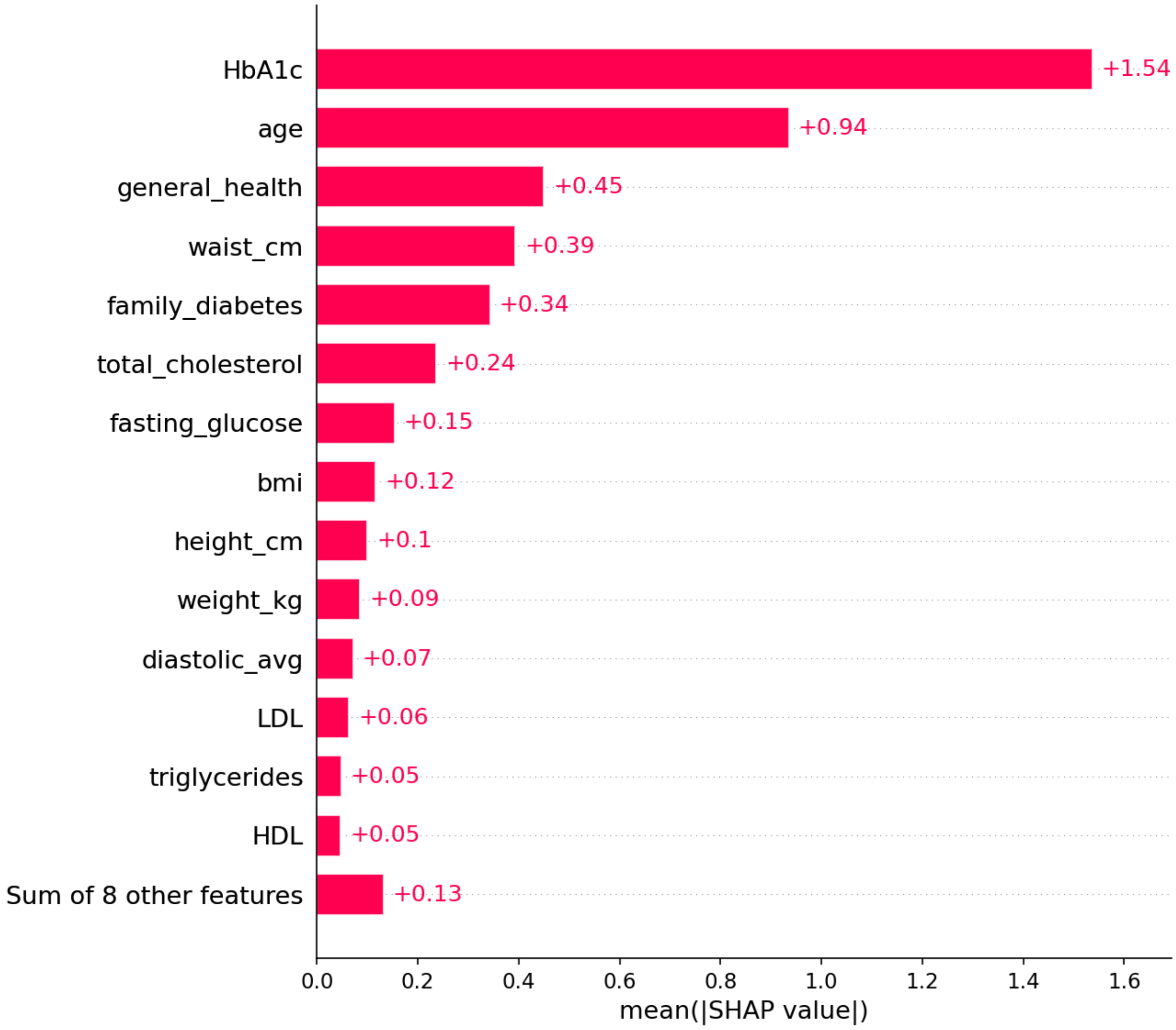
The most important features for whole people.



Furthermore, a feature importance bar plot summarizes the relative importance of each variable, offering an overview of the most influential factors in diabetes risk prediction.

Global Explanation / 模型整體解釋

The most important features for whole people.



8. Limitations and Future Work

(1) USA Dataset v.s. Taiwan Dataset

This project is built using NHANES, a U.S.-based dataset, and differences in genetics, lifestyle, and healthcare systems may limit the direct generalizability of risk estimates to the Taiwanese population.

However, the proposed modeling pipeline and cloud architecture are **population-agnostic**, and the primary goal is to demonstrate a reproducible and scalable framework rather than clinically calibrated risk scores.

Future work includes adapting the system to Taiwan-specific health datasets to improve population relevance and calibration, as well as incorporating longitudinal data and wearable sensor information.

(2) Other Diseases

The modular system design naturally supports extension to **other chronic diseases**, such as cardiovascular disease or metabolic disorders, by retraining disease-specific models while reusing the same data pipeline, explanation module, and cloud infrastructure.