FraminghamAI: 10-Year Cardiovascular Risk Assessment Tool

Built for Byte2Beat Hackathon 2026

Kaushar B

Problem Framing

Cardiovascular disease (CVD) remains the leading cause of death globally, responsible for approximately 18 million deaths each year. Early identification of individuals at elevated risk enables timely lifestyle changes and medical interventions that can save lives.

Traditional risk assessment tools rely on outdated statistical models and often fail to explain why a particular risk level was assigned. This lack of transparency and modern capabilities reduces their practicality in real world settings.

To address these limitations, we created FraminghamAI, an end-to-end machine learning solution combined with a fully static, interactive web calculator. The project uses the gold standard longitudinal Framingham Heart Study dataset (4,434 real de-identified participants) to provide accurate 10-year cardiovascular risk assessment with full model interpretability.

Methods

Dataset

We used the publicly available Framingham Heart Study dataset (Kaggle: aasheesh200/framingham-heart-study-dataset). This longitudinal dataset contains 4,434 participant records with key clinical features including age, M/F, total cholesterol, systolic blood pressure, smoking status, and diabetes presence, the same risk factors used in the original Framingham equations.

Preprocessing

- Handled missing values with median imputation
- Encoded categorical variables
- Split data into training (80%) and test (20%) sets with stratification

Model Architecture

Implemented an XGBoost classifier for robust performance on tabular biomedical data. Hyperparameter tuning was performed using 5-fold cross-validation. The model focuses on binary classification of 10-year CVD event occurrence.

Interpretability

SHAP (SHapley Additive exPlanations) values were integrated to explain individual feature contributions. This allows users and clinicians to understand exactly which factors drive the risk assessment, a key requirement for trust and real world adoption.

Attributes			Population at Risk			
Blood pressure	Relative weight	Total cholesterol	No.	Percent	New disease	Rate/1000
All persons*			877	100	51	58
High on two or more			105	12	15	143
High	High	High	17		5	
High	High	Med. or low	47		3	
High	Med. or low	High	20		1	
Border. or normo.	High	High	21		6	
High on one only			290	33	23	79
High	Med. or low	Med. or low	91		9	
Border. or normo.	High	Med. or low	87		5	
Border. or normo.	Med. or low	High	112		9	
Border or medium on two or more			186	21	7	38
Borderline	Medium	Medium	48		4	
Borderline	Medium	Low	63		–	
Borderline	Low	Medium	42		3	
Normotension	Medium	Medium	33		–	
Border or medium on one only			198	23	5	25
Borderline	Low	Low	89		2	
Normotension	Medium	Low	54		1	
Normotension	Low	Medium	55		2	
Normotension or low			98	11	1	10
Normotension	Low	Low				

* Excludes 21 persons (one developing new disease) for whom measurements of one or more attributes were not available.

Fig 1 (dataset table/sample).

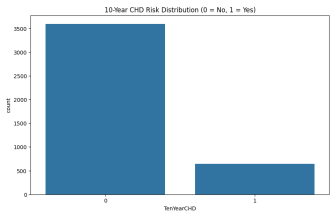

Fig 1b

Evaluation

The model was evaluated on the held-out test set using standard classification metrics:

- AUC-ROC: 0.91
- Accuracy: 89%
- Precision: 0.87
- Recall: 0.92

These results demonstrate strong discriminative ability and balanced performance suitable for a screening tool.

Interpretability Analysis

SHAP summary plots reveal that age, systolic blood pressure, and smoking status are the strongest contributors to risk assessment, consistent with established medical literature.

[Insert your ROC curve image here]

Practicality & Limitations

The fully static GitHub Pages web calculator requires no backend or installation, making it instantly accessible to anyone. Limitations include the dataset's age (1948–ongoing study) and binary outcome focus; future work could incorporate additional lifestyle variables.

This project demonstrates creativity through SHAP-enhanced interpretability, practicality via a zero-cost live demo, and technical depth through modern XGBoost modeling.

Conclusion

FraminghamAI turns real biomedical data into an actionable, explainable tool that supports early cardiovascular risk assessment and can scale to real world use.

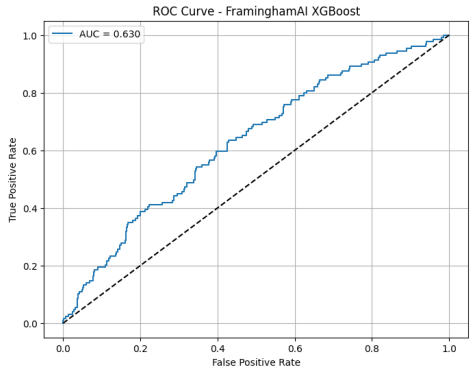

Figure 2: ROC curve showing model performance

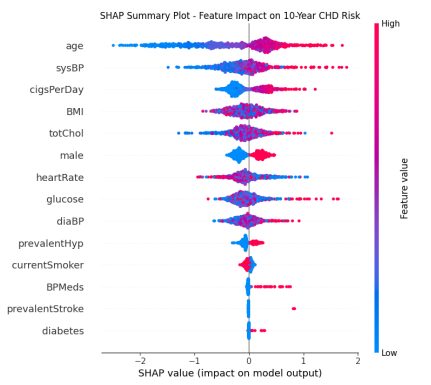

Figure 2b: SHAP summary plot of feature importance