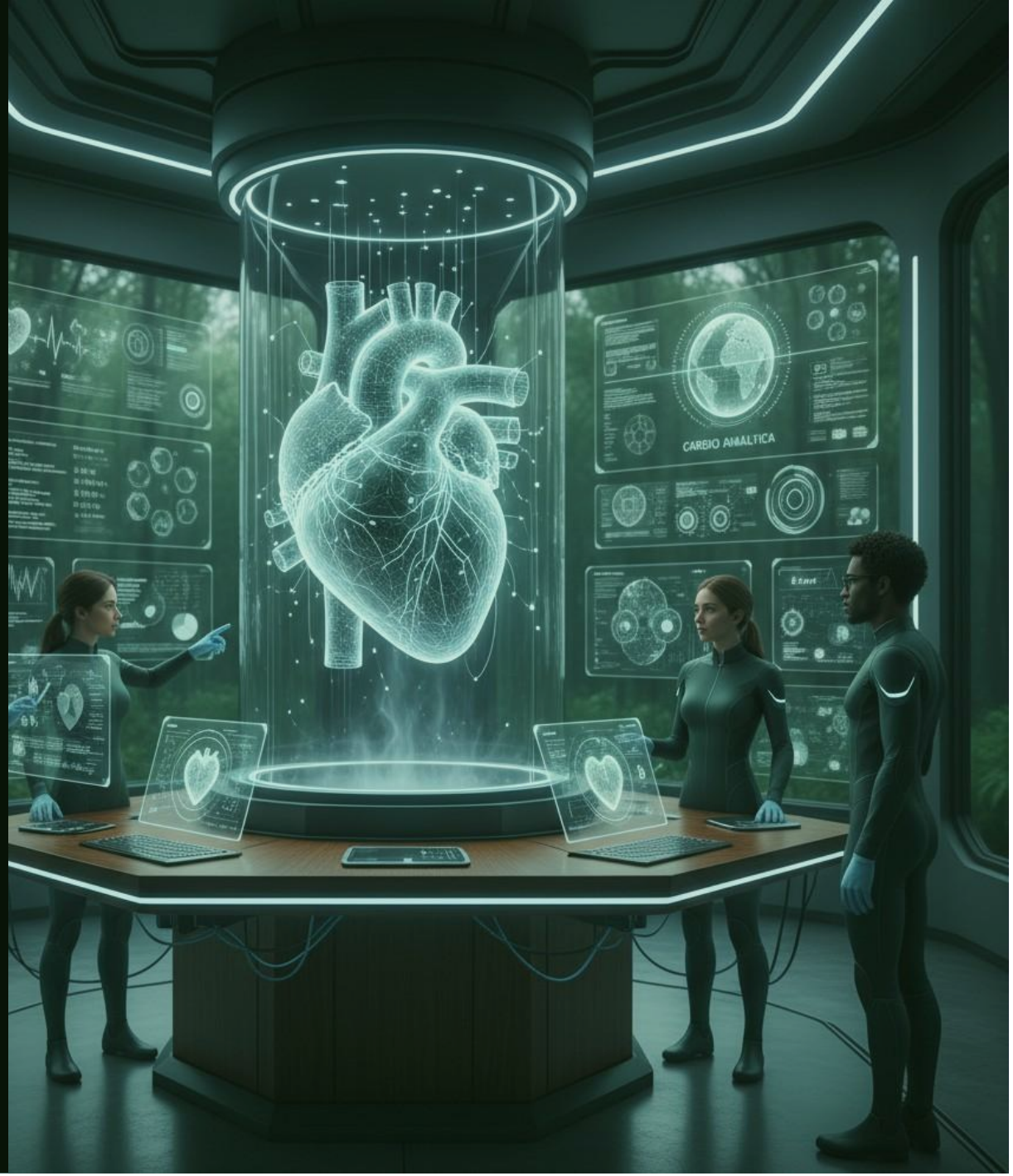
Predicting Cardiovascular Disease (CVD) from Lifestyle Behaviors

This project explores the feasibility of predicting heart disease risk using solely non-clinical behavioral, demographic, and lifestyle data.

Team B08 : Zixiao Jiao, Junyi Liu, Hanchao Tang, Zixuan Zhu



Problem Definition

Research Question:

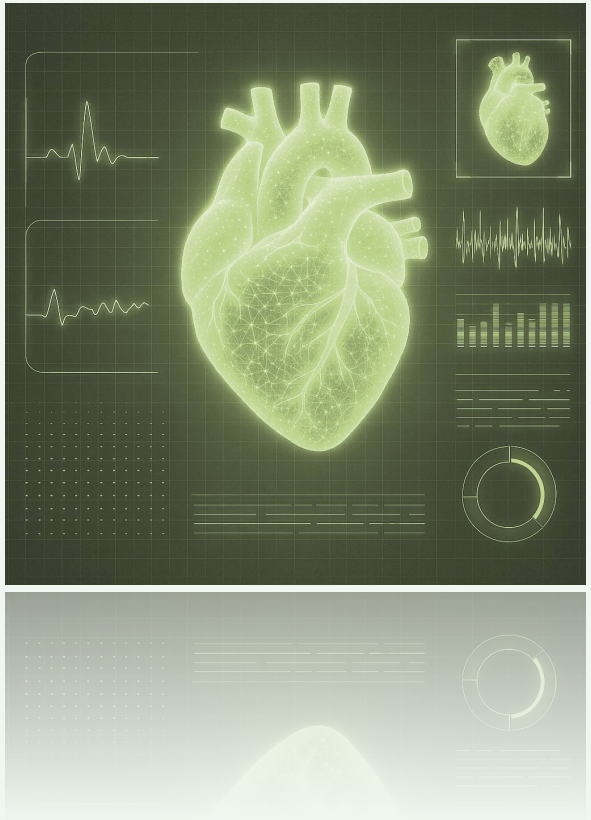
Can we classify whether an adult has heart disease **using only demographic, behavioral, and self-reported health data**, while **excluding** all clinical test results or medical diagnoses?

THE APPROACH:

We test whether **non-clinical survey data alone** can predict heart disease risk in a realistic, resource-limited scenario.

THE GOAL:

To test the predictive power of lifestyle data alone, simulating real-world constraints where medical screening is unavailable.



Why This Framing Matters

Success Scenario

If models built exclusively on behavioral and lifestyle data demonstrate strong predictive performance (particularly high recall), it validates the feasibility of creating widespread, low-cost early warning systems for cardiovascular risk.

Failure Scenario

Conversely, if model accuracy collapses without the inclusion of clinical variables, it would strongly suggest that behavioral interventions are iinsufficient on their own, reinforcing the indispensable role of clinical screening.

Why This Problem Matters

The Paradigm Shift: From Reactive to Proactive

Predicting heart disease using behavioral data creates major public-health value. Simple survey responses can help shift healthcare from costly treatment to early, low-cost prevention.

Public Health Agencies

- Identify high-risk regions **without large-scale clinical testing**
- Support targeted outreach and education

Hospitals & Insurers

- Stratify populations based on behavioral risk
- Allocate screening resources more efficiently
- Reduce long-term medical costs

Policy Makers

- Evidence supports **policies promoting healthy behaviors**
- Healthier populations reduce chronic-disease burden

Individuals

- Personalized lifestyle-based feedback
- Earlier awareness → earlier action
- No need for expensive clinical tests

Data Source Overview

CDC BRFSS 2015

Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset —-U.S. Centers for Disease Control and Prevention (CDC)

Comprehensive, **large-scale health survey** conducted via telephone interviews with over 400,000 respondents annually.

Nationally Representative

The survey sampling **methodology** ensures that the findings can be extrapolated to the broader U.S. Population relevant for **national public health** policy

Standardized Collection

Collect Data from **telephone interviews** provide consistent format for responses **from dietary habits and physical activity to smoking status**

Behavioral Risk Focus

capture the **behavioral and environmental factors** associated with **chronic diseases**

👉aligns with our goal of **simulating non-clinical prediction scenarios**.

Target: CVDCRHD4

"(Ever told) you had coronary heart disease or myocardial iinfarction?"

Yes (1) or No (2) responses for clean binary classification

Dataset Technical Description

▼ DATA ENGINEERING FUNNEL

RAW BRFSS DATA

~330,000

Respondents

↓

TARGET DEFINED

253,795

Valid Yes/No responses

↓

FINAL SAMPLE

100,000

Rows for Modeling

Balanced for performance & feasibility

Feature Composition

After a rigorous **cleaning and feature selection** process, we finalized a set of 17 predictors

- Mostly Categorical
- Numeric-coded survey answers

18

TOTAL COLUMNS

(17 Predictors + 1 Target)

Target Distribution

The dataset exhibits a highly imbalanced target class. This is a **critical modeling** challenge, as standard algorithms tend to favor the majority class.

Majority Class (Negative)

Positive Class (Heart Disease)

Requires mitigation strategies like class weighting and focusing on recall.

Feature Selection Strategy

STRATEGY:
To simulate a **realistic, non-clinical** screening scenario removed features that could leak information about a respondent's existing health status

Input: The Feature Groups

- 1. Demographics
- 2. Lifestyle
- 3. Health Perception
- 4. Access to Care — LEAKAGE RISK

Preventing "Data Leakage"

A critical challenge

—— predictors act as proxies for the outcome (e.g., blood pressure medication implies heart disease)

Dataset Version: Original — Baseline

The full set of cleaned features, including potential leakage sources.

Dataset Version: Strong Reduced — Strict

Removed direct indicators of other chronic diseases but kept variables related to healthcare access.

Dataset Version: All Reduced ★ — FINAL MODEL

Removed all chronic disease indicators AND signals related to healthcare access.

This represents the most stringent and realistic test of a purely behavior-based prediction model.

Data Cleaning Summary

process :

numeric-coded categorical fields

inconsistent special values

missing data

👉ensure **model robustness** and **prevent misleading interpretations**

Numeric-coded Categories

Action Taken:
All fields were carefully recorded and grouped based on official codebooks. Income ranges mapped to ordered categories.

REASON: Prevents false ordinal assumptions (e.g., 5 > 4 for nominal data).

Special Code Meanings

Action Taken:
Special codes (e.g., 7 for 'Don't know', 9 for 'Refused') were mapped into clear, distinct buckets like 'Unknown'.

REASON: Removes noise; allows model to treat non-responses as specific info.

High Missing Data (>50%)

Action Taken:

The variables `ADSLEEP` and `POORHLTH` were removed from the dataset entirely.

REASON: Imputing >50% missing data is unreliable and introduces bias.

Complex Range Encoding

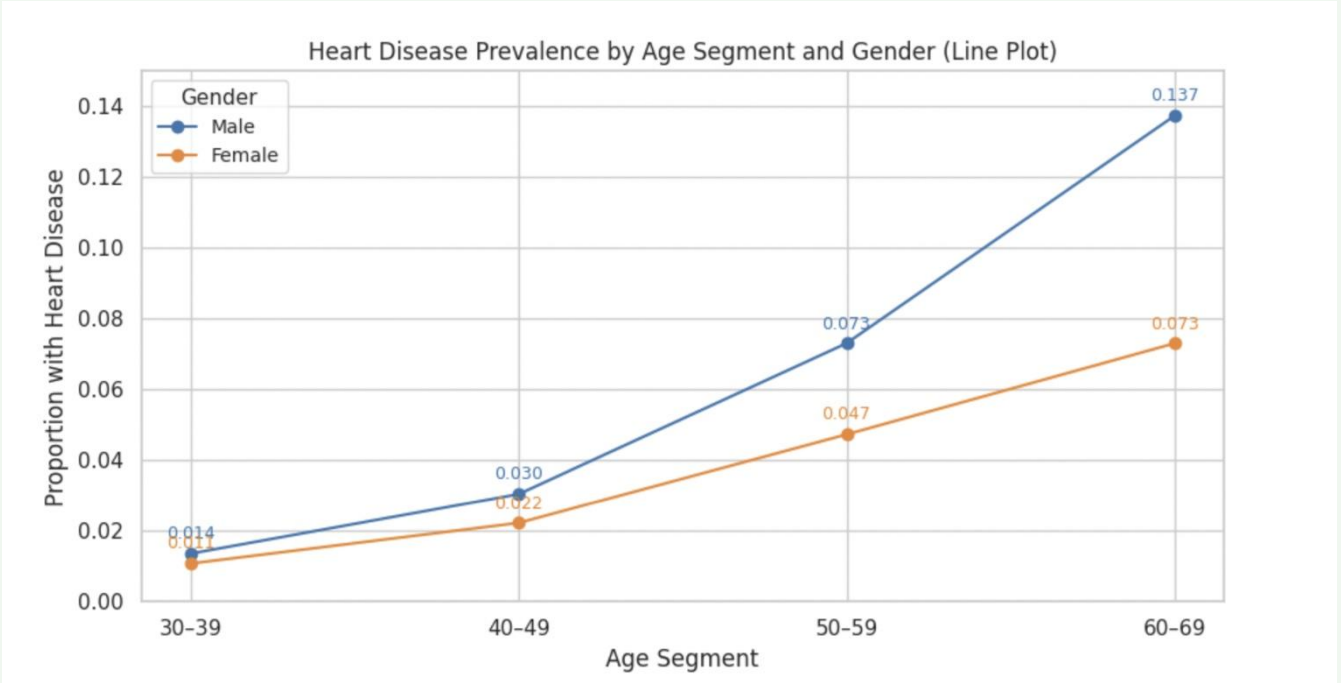
Action Taken:
Collapsed original encodings (e.g., 100s for daily) into simpler categorical buckets like 'Daily', 'Weekly', and 'Monthly'.

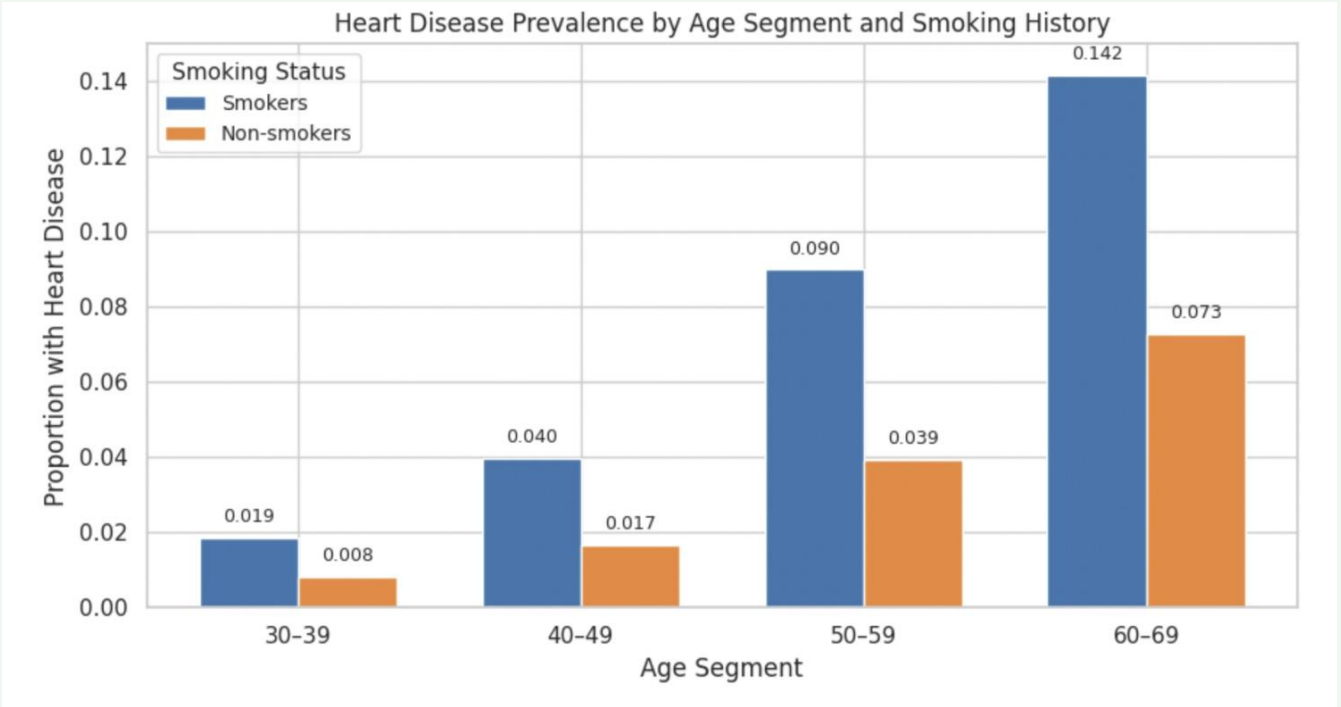
REASON: Simplifies feature space; improves generalization on frequency patterns.

Exploratory Data Insights

CORE INSIGHT

- All features show **statistically significant** links to heart disease.
- **Age** and **smoking history** are strong predictors.
- **Male prevalence** is consistently higher across age groups.





EDA Insights

CORE INSIGHT

- Chi Square analysis between **features** and **target variable**
- Every features shows a statistically significant association with heart disease
- Features that have leakage risks have more significant p-value since they have stronger association with target variable

Group	Predictor	p-value	Notes
1. Demographics			
	_AGEG5YR	0.000000e+00	Strong age risk factor
	SEX	2.54e-78	Male higher risk
	MARITAL	3.33e-204	Social support indicator
	INCOME2	8.04e-108	Socioeconomic status
	EDUCA	5.19e-43	Education level
2. Lifestyle			
	_SMOKER3	7.89e-210	Smoking status
	SMOKE100	2.97e-161	Lifetime smoking
	_TOTINDA	1.76e-78	Physical activity
	ALCDAY5	1.83e-85	Alcohol consumption
	FRUIT1	5.17e-13	Diet (fruit)
	VEGETAB1	4.48e-05	Diet (vegetables)
3. Health Perception			
	GENHLTH	0.000000e+00	General health
	PHYSHLTH	0.000000e+00	Physical health days
	MENTHLTH	3.08e-47	Mental health days
4. Access to Care (LEAKAGE RISK)			
	PERSDOC2	3.48e-173	Has personal doctor
	CHECKUP1	1.00e-110	Recent checkup
	CHOLCHK	1.26e-291	Cholesterol check
	HLTHPLN1	1.59e-30	Insurance
	MEDCOST	3.29e-04	Cost barrier
5. Clinical Conditions (LEAKAGE RISK)			
	DIABETE3	0.000000e+00	Diabetes
	BPHIGH4	0.000000e+00	High blood pressure
	BLOODCHO	5.15e-123	High cholesterol
	CHCKIDNY	0.000000e+00	Kidney disease
	CVDSTRK3	0.000000e+00	Stroke

Modeling Approach

PORTFOLIO STRATEGY

We tested ML models to see how well our data predicts outcomes.

CRITICAL METRIC

Prioritizing Recall Over Accuracy

In a clinical context, a false negative (missing a person with heart disease) is a far more dangerous error than a false positive.

"Our primary goal was... to correctly identify true positive cases."

Logistic Regression

Included as an interpretable baseline model. Its coefficients provide clear insights into the influence of each feature, valuable for explaining the 'why'.

KNN

Tested to explore distance-based methods. Performed poorly due to the 'curse of dimensionality' in our wide feature space.

XGBoost

Included as a state-of-the-art benchmark. Performed poorly, likely struggling with the highly categorical data without advanced engineering.

Decision Tree

Chosen for its ability to handle high-dimensional, sparse space from one-hot encoding. Not sensitive to distance-based issues.

Random Forest

A powerful ensemble method that reduces variance. Known for strong performance and high recall in similar classification tasks.

Voting & Stacking Ensembles

Built to combine predictions of multiple models, blending characteristics to optimize bias-variance trade-offs and learn complex decision boundaries.

Logistic Regression and limitation

METHODOLOGY

Logistic Regression

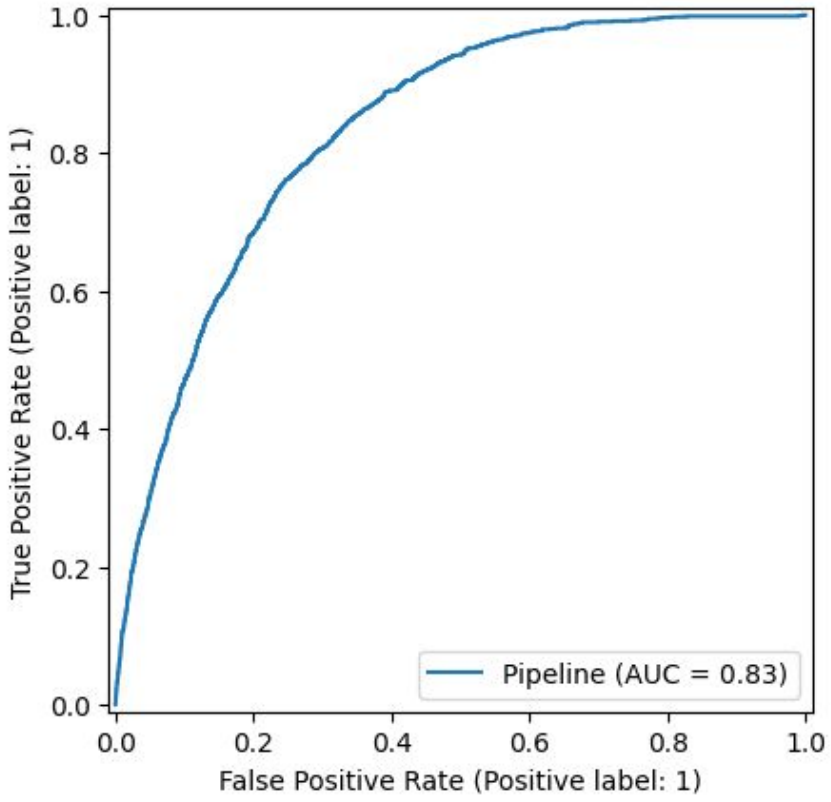
PRIMARY GOAL: Accuracy

- Very small positive class (only ~5%) means the threshold of 0.5 is inappropriate.
- Severe class imbalance skews the model toward predicting the majority class (No heart disease).
- AUC = 0.83 shows overall good performance

```
Accuracy: 0.942
Balanced accuracy: 0.506
              precision    recall  f1-score   support

           0      0.943     0.999     0.970     18847
           1      0.519     0.012     0.024      1153

    accuracy                         0.942     20000
   macro avg      0.731     0.506     0.497     20000
weighted avg      0.919     0.942     0.916     20000
```



orig_feature	coefficient
GENHLTH	2.17613
_AGEG5YR	1.74588
SMOKE100	0.759304
EDUCA	0.742582
EMPLOY1	0.726495
SEX	0.671692
ALCDAY5	0.565724
PERSDOC2	0.469723
MARITAL	0.2959
MENTHLTH	0.276687
VEGETAB1	0.270678
PHYSHLTH	0.257443
_SMOKER3	0.238316
INCOME2	0.168644
FRUIT1	0.126852
_TOTINDA	0.029464

Hyperparameter Optimization Framework

METHODOLOGY

Systematic Tuning for Sensitivity

We employed **RandomizedSearchCV** to systematically tune each model.

PRIMARY GOAL: RECALL

Optimized for 'recall'

clinical goal: maximizing the detection of positive heart disease cases.

AVOIDS

False Negatives (Critical Misses)

Example Tuning Search Spaces

Logistic Regression

- C
- Penalty
- Class_Weight
- Max_iteration

K-Nearest Neighbors (KNN)

- N_neighbors
- Weights
- P

Random Forest

- N_estimators
- Depth
- Leaf size
- Split size
- Feature sampling
- Class weighting

XGBoost

- N_estimators
- Max_depth
- Learning_rate

Decision Tree & Random Forest— Insights & Limitations

Decision Tree

- Achieves the **highest recall (0.78)** and captures nonlinear lifestyle risk patterns.
- Shallow structure prevents overfitting and keeps results interpretable.
- Best choice when **interpretability** is required.

Random Forest

- More **stable and robust** with 271 trees and depth 5.
- Best for **prediction reliability** over interpretability.

Rank	Mean Test Score	Class Weight	Min Samples Split	Min Samples Leaf	Max Features
1	0.768085	balanced	38	3	None
2	0.766157	balanced	24	16	None
3	0.730105	balanced	5	10	sqrt
4	0.727961	balanced	13	1	sqrt
5	0.724951	balanced	28	17	None

Rank	Mean Test Score	n_estimators	max_depth	min_samples_leaf	min_samples_split	max_features	class_weight
1	0.783091	271	5	5	36	None	balanced
2	0.756702	314	8	16	16	log2	balanced
3	0.755201	149	7	17	37	log2	balanced_subsample
4	0.754557	330	7	5	43	sqrt	balanced_subsample
5	0.750267	230	7	3	38	sqrt	balanced

KNN Clustering–Insights & Limitations

KNN – Key Insights

- **Very low recall (~0.07)** → cannot identify heart-disease cases.
- Categorical one-hot features make **distance metrics ineffective**.
- **Low Positive Cases**

Rank	Mean Test Recall	n_neighbors	weights	p (distance metric)
1	0.07788	2	distance	2
1	0.07788	2	distance	2
3	0.068223	3	uniform	2
4	0.06758	3	uniform	1
5	0.024029	7	uniform	2

XGBoost – Key Insights

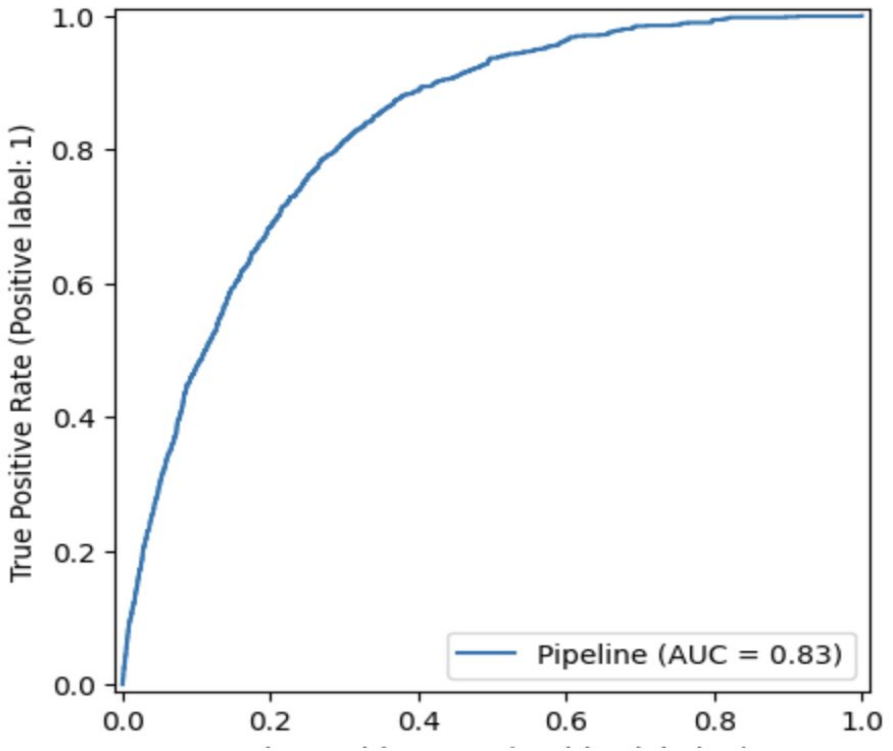
XGBoost

- Achieves **moderate recall (~0.18–0.20)**
- Handles nonlinear patterns well, but **lifestyle survey features (categorical + one-hot)** reduce its boosting advantage.
- Model becomes **complex and harder to interpret**

```
=== XGBoost ===
Accuracy: 0.929
Balanced Accuracy: 0.577
              precision    recall  f1-score   support

           0      0.951     0.975     0.963     18847
           1      0.302     0.179     0.225      1153

    accuracy                         0.929     20000
   macro avg      0.627     0.577     0.594     20000
weighted avg      0.914     0.929     0.920     20000
```



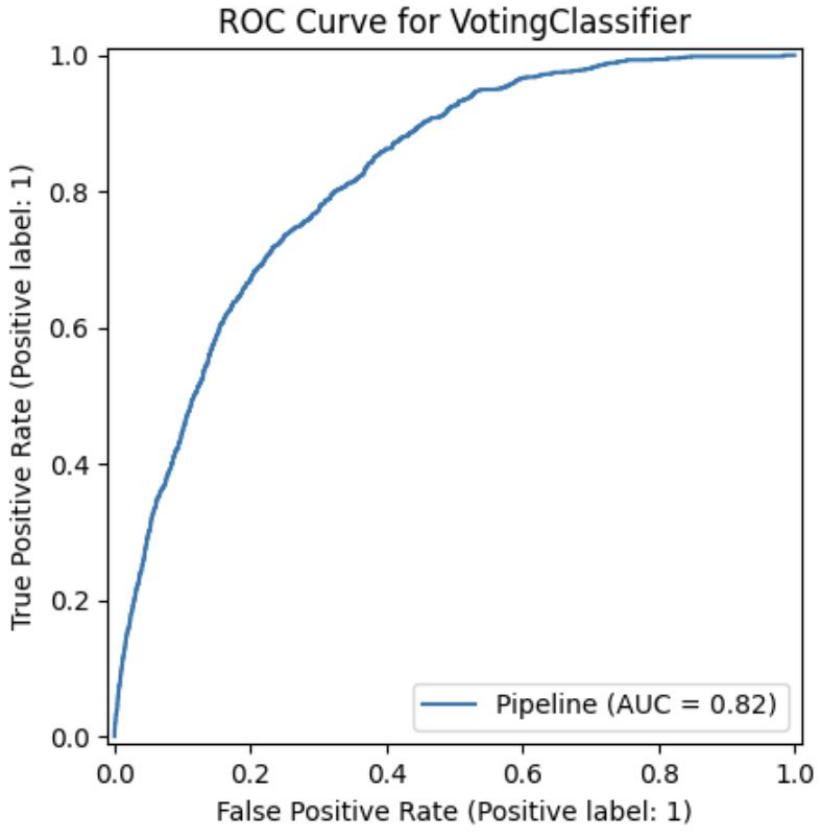
Voting & Stacking

```
=== VotingClassifier (Soft Voting) ===
Accuracy: 0.720
Balanced accuracy: 0.737
              precision    recall  f1-score   support

           0      0.980     0.718     0.829     18847
           1      0.141     0.757     0.238      1153

    accuracy                         0.720     20000
   macro avg      0.560     0.737     0.533     20000
weighted avg      0.931     0.720     0.794     20000

Recall score for Voting Classifier: 0.757
```

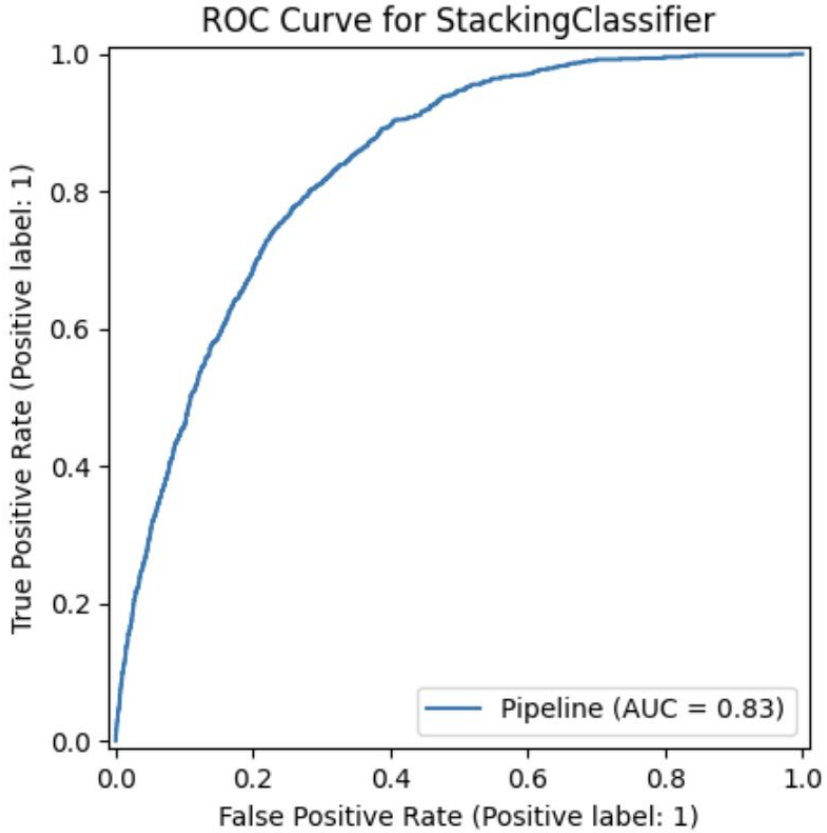



```
=== StackingClassifier ===
Accuracy: 0.741
Balanced accuracy: 0.759
              precision    recall  f1-score   support

           0      0.982     0.738     0.843     18847
           1      0.154     0.779     0.257      1153

    accuracy                         0.741     20000
   macro avg      0.568     0.759     0.550     20000
weighted avg      0.934     0.741     0.809     20000

Recall score for Stacking Classifier: 0.779
```



Challenge & Resolution

CHALLENGE	THE DIFFICULTY	STRATEGIC RESOLUTION
Missing categorical data	Cannot mean-impute	Imputed symbolic "Unknown" to prevent leakage
High Dimensionality One-Hot space	Distance metrics break down	Switched to trees & ensembles optimized for recall
Invalid Coded values(777,999 etc.)	Survey coding inconsistent	Recategorized using BRFSS documentation
Leakage from medical history	Inflated early model performance	Created behavioral-only dataset version
Class Imbalance	Few positive cases hurt sensitivity	Used class-weight balancing + recall evaluation

Conclusion: Model's Inability & Future Work

Further Development

Richer Behavioral Signals

- Intensity of smoking (pack-years), diet quality score, sleep regularity
- Not just whether they smoke/exercise, but how much & how often

Environmental & Contextual Indicators

- Air pollution index, food desert score, walkability + green space availability
- Stress indicators from socioeconomic + geographic context.

Psychological & Social Factors

- Stress, depression scale, loneliness, social support level
- Behavioral disease risk is often mediated by mental health

MODEL's Inability

Limited AUC ceiling (~0.83 across all models)

- Feature set may not fully capture clinical drivers of heart disease, placing an upper bound on predictive performance.

Complex ensemble models fail to outperform simple logistic regression

- Voting/stacking add complexity without additional signal, suggesting limited nonlinear structure in the dataset.

Limited interpretability of advanced models

- Hard to assess how individual features influence predictions, reducing clinical trust and decision transparency.

Trade-off between positive and negative accuracy

- Improving recall for positive cases increases false positives and reduces true negatives, forcing difficult threshold decisions.

Colab notebook link:https://colab.research.google.com/drive/1Ze06OIQrFnRRyPAFju894YF4r0ysL7JE?usp=sharing