# AI-Driven Early Detection of Chronic Diseases Using Lifestyle and Behavioral Data from CDC BRFSS 2024

## Problem Statement

Chronic diseases such as diabetes, heart disease, asthma, and arthritis often progress silently before symptoms appear. Late detection leads to serious complications and higher healthcare costs. This project aims to develop an AI-powered predictive model that identifies individuals at risk of developing chronic diseases early, based on lifestyle and demographic factors. The goal is to empower healthcare providers with a data-driven screening tool that flags high-risk patients for preventive testing and lifestyle interventions.

## Team Details

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Responsibility** |
| Sibangi Subhadarsani | Data Analyst | Model design and Streamlit app integration |
| Bijaya Ram Shrestha | Data Analyst | Data cleaning, Model design and Streamlit app integration |
| Abhishek Joshi | Web Engineer + Data Analyst | Streamlit UI/UX design and deployment, Data verification and exploratory analysis |

## Dataset

The project utilizes the CDC Behavioral Risk Factor Surveillance System (BRFSS 2024) dataset, a large-scale public health dataset containing demographic, behavioral, and health indicators from U.S. adults. The dataset enables population-level prediction of chronic disease risks.

### Column Selections

The selected features and target variables are as follows:

#### Feature Variables

| **Variable** | **Full Form / Meaning** | **Type** | **Description / Relevance** |
| --- | --- | --- | --- |
| **PHYS14D** | Physical Health – 14 Days | Numeric (Days) | Number of days during the past 30 days when the respondent’s physical health was not good. Higher values may indicate chronic illness or poor lifestyle. |
| **TOTINDA** | Physical Activity (Total) | Categorical (Yes/No) | Indicates whether the person has done any physical activity or exercise (like running, gardening, or walking) in the past 30 days. Lack of activity is a strong predictor of obesity and chronic disease. |
| **SEX** | Biological Sex | Binary (1=Male, 2=Female) | Helps identify gender-related differences in disease risk; for instance, men are more likely to have heart attacks at a younger age. |
| **AGE\_G** | Age Group | Categorical | Age bracket of respondent. Age is a key predictor of chronic diseases such as diabetes and cardiovascular conditions. |
| **BMI5CAT** | Body Mass Index Category | Categorical (1–4) | Derived from BMI; classifies individuals as underweight, normal, overweight, or obese. Obesity strongly correlates with both diabetes and heart disease. |
| **EDUCAG** | Education Level | Categorical | Reflects the highest education attained. Lower education is often linked to poorer health literacy and higher disease risk. |
| **INCOMG1** | Income Level | Categorical | Indicates annual household income category. Lower income is associated with limited access to healthcare and higher chronic disease prevalence. |
| **RFSMOK3** | Smoking Status (Risk Factor) | Categorical (Yes/No) | Identifies whether the person currently smokes or has ever smoked 100+ cigarettes. Smoking increases cardiovascular and respiratory risks. |
| **DRNKANY6** | Alcohol Consumption | Categorical (Yes/No) | Indicates whether the respondent has consumed alcohol in the past 30 days. Excessive drinking raises the risk for diabetes and heart disease. |
| **SSBSUGR2** | Sugar-Sweetened Beverage Intake | Ordinal (Low/Med/High) | Measures frequency of consumption of sugary drinks. High intake correlates with obesity and insulin resistance. |

Target Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Full Form / Meaning** | **Type** | **Description / Relevance** |
| **MICHD** | Myocardial Infarction / Coronary Heart Disease | Binary (Target Variable 1) | Indicates whether the person has ever been told they had a heart attack, coronary heart disease, or angina. |
| **DIABETE4** | Diabetes Status | Binary (Target Variable 2) | Indicates whether the respondent has been diagnosed with diabetes (excluding gestational diabetes). Serves as a key chronic disease indicator. |

## Tools and Technologies

* Programming: Python 3.10
* Libraries: Pandas, NumPy, Scikit-learn, XGBoost, Seaborn, Matplotlib, Imbalanced-learn, Joblib
* Visualization & Interface: Streamlit
* Version Control: Git & GitHub
* Platform: VS Code

## Methodology

1. Data Cleaning and Preprocessing: Missing values handled, categorical encoding applied, and features scaled using StandardScaler.
2. Exploratory Data Analysis: Examined correlations between lifestyle factors and disease prevalence.
3. Model Development: Two separate models were developed—one for heart disease (MICHD) and another for diabetes (DIABETE4). Baseline logistic regression models were first implemented, followed by enhanced models using XGBoost for performance improvement.
4. Stacked Ensemble: Combined XGBoost and Logistic Regression using StackingClassifier to achieve higher ROC-AUC scores for heart disease model.
5. Model Evaluation: Metrics such as Accuracy, F1-Score, and ROC-AUC were used. ROC and Precision-Recall curves were visualized.
6. Deployment: Trained models and scalers were saved using Joblib and deployed on Streamlit for interactive web-based prediction.

## Results

The logistic regression model provided a strong baseline, while XGBoost improved predictive accuracy and recall across all target diseases. The ROC-AUC scores demonstrated reliable separability between positive and negative classes, confirming model robustness. Stacked ensemble models achieved the best trade-off between interpretability and accuracy.

Model 1: Heart Disease Detection result

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score** | **TN** | **FP** | **FN** | **TP** |
| **Stacked (XGB → LR)** | 0.852 | 0.317 | 20,322 | 2,209 | 1,474 | 853 |
| **Logistic Regression** | 0.857 | 0.318 | 20,476 | 2,055 | 1,499 | 828 |
| **Decision Tree** | 0.850 | 0.227 | 20,591 | 1,940 | 1,782 | 545 |

Stacked model selected for the prediction model for the following reasons:

1. Captures complex feature interactions, non-linear relationships, and threshold effects.
2. Generalize better on new unseen data.
3. Avoid underfitting (especially if relationships are nonlinear or interaction-based).
4. Detects 25 more true disease cases, which is crucial for *medical or chronic disease prediction*

Model 2: Diabetes Detection result

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score (Weighted Avg.)** | **Test ROC AUC** |
| **Logistic Regression** | 0.8315 | 0.7721 | 0.7694 |
| **Random Forest** | 0.8314 | 0.7754 | 0.7003 |
| **HistGradientBoosting** | 0.8318 | 0.7634 | 0.7736 |
| **GradientBoosting** | 0.8318 | 0.7690 | 0.7757 |

## Challenges and Future Improvements

Challenges:

* Handling data imbalance and missing survey responses.
* Translating survey indicators into clinically meaningful variables.
* Ensuring fairness and avoiding bias across demographic groups.

Future Enhancements

* Integrate real-time wearable sensor data for better lifestyle tracking.
* Incorporate Deep Learning architectures for non-linear pattern recognition.
* Use Explainable AI (XAI) for transparent predictions.
* Extend dashboard to include clinical alerts and EHR data integration.

## Conclusion

This project demonstrates how AI and public health datasets can be leveraged for preventive healthcare. By combining machine learning algorithms with behavioral and lifestyle data, the predictive model assists healthcare providers in identifying high-risk patients early and promoting proactive medical interventions. The Streamlit prototype showcases the potential for scalable, AI-driven health screening systems.