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

1. 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.

2. Team Details

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

3. 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.

3.1 Column Selections

The selected features and target variables are as follows:

3.1.1 Feature Variables

Variable	Full Form / Meani	Type	Description / Relevance
PHYS14D	Physical Health – Days	Numeric (Days)	Number of days during the past 30 days when the respondent's physical health was not good. Higher v may indicate chronic illness or poor lifestyle.
TOTINDA	Physical Activity (T	Categorical (Yes/No	Indicates whether the person has done any physical activity or exercise (like running, gardening, or walki the past 30 days. Lack of activity is a strong predicto obesity and chronic disease.
SEX	Biological Sex	Binary (1=Male, 2=Female)	Helps identify gender-related differences in disease for instance, men are more likely to have heart attac a younger age.

AGE_G	Age Group	Categorical	Age bracket of respondent. Age is a key predictor of chronic diseases such as diabetes and cardiovascula conditions.
BMI5CAT	Body Mass Index Category	Categorical (1–4)	Derived from BMI; classifies individuals as underweig normal, overweight, or obese. Obesity strongly corre with both diabetes and heart disease.
EDUCAG	Education Level	Categorical	Reflects the highest education attained. Lower educa s often linked to poorer health literacy and higher dis isk.
NCOMG1	ncome Level	Categorical	ndicates annual household income category. Lower ncome is associated with limited access to healthca and higher chronic disease prevalence.
RFSMOK3	Smoking Status (R Factor)	Categorical (Yes/No	dentifies whether the person currently smokes or ha ever smoked 100+ cigarettes. Smoking increases cardiovascular and respiratory risks.
DRNKANY6	Alcohol Consumpt	Categorical (Yes/No	ndicates whether the respondent has consumed alc n the past 30 days. Excessive drinking raises the ris diabetes and heart disease.
SSBSUGR2	Sugar-Sweetened Beverage Intake	Ordinal (Low/Med/H	Measures frequency of consumption of sugary drinks High intake correlates with obesity and insulin resista

Target Variables

Variable	Full Form / Meanir	Type	Description / Relevance
MICHD	Myocardial Infarctio Coronary Heart Dise	Binary (Target Var 1)	Indicates whether the person has ever been told the a heart attack, coronary heart disease, or angina
DIABETE	Diabetes Status	Binary (Target Var 2)	Indicates whether the respondent has been diagno with diabetes (excluding gestational diabetes). Serve a key chronic disease indicator.

4. Tools and Technologies

Ø Programming: Python 3.10

Ø Libraries: Pandas, NumPy, Scikit-learn((LogisticRegression, RandomForestClassifier, HistGradientBoostingClassifier, GradientBoostingClassifier),, XGBoost, Seaborn, Matplotlib, Imbalanced-learn, Joblib

Ø Visualization & Interface: Streamlit

Ø Version Control: Git & GitHub

Ø Platform: VS Code

5. Methodology

1. Data Cleaning and Preprocessing: Missing values handled, categorical encoding applied, and features scaled using StandardScaler.

The initial dataset (LLCP2024.XPT) was filtered to include only completed interviews (DISPCODE == 1100).

Missing values were handled by first replacing specific numerical codes (e.g., 7, 9, 777, 999) with NaN for 11 key features, and then dropping all rows containing any NaN values using .dropna(). This resulted in a cleaned dataset of 82,859 records.

The SSBSUGR2 feature was binned into three categories ('Low', 'Medium', 'High') and then converted to numerical format using OrdinalEncoder.

All features were scaled using StandardScaler before model training.

2. Exploratory Data Analysis: Examined correlations between lifestyle factors and disease prevalence. A subset of 14 relevant columns was selected for the analysis. Initial exploration included examining data types, value counts for target variables (DIABETE4 and DRDXAR2), and assessing the extent of missing values.

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. The notebook code focuses on developing a model for diabetes (DIABETE4). Four baseline models were implemented and compared: LogisticRegression, RandomForestClassifier, HistGradientBoostingClassifier, and GradientBoostingClassifier. The HistGradientBoostingClassifier was selected for performance improvement, and its hyperparameters were tuned using RandomizedSearchCV. The best parameters found were {'max_leaf_nodes': 70, 'max_iter': 300, 'learning_rate': 0.01}.

4. Combined XGBoost and Logistic Regression using StackingClassifier to achieve higher ROC-AUC scores for heart disease model and the final model used for diabetes disease was the tuned HistGradientBoostingClassifier.

5. Model Evaluation: Metrics such as Accuracy, F1-Score, and ROC-AUC were used. ROC and Precision-Recall curves were visualized. The performance of the final model was evaluated using several metrics. The final model achieved a test accuracy of 83.16%.A classification_report was generated to evaluate these metrics for each class. The final model achieved a macro-averaged ROC-AUC score of 0.7471 on the test set for diabetes.

6. Deployment: Trained models and scalers were saved using Joblib and deployed on Streamlit for interactive web-based prediction. The trained HistGradientBoostingClassifier model and the StandardScaler object were serialized and saved into separate files (hist_model_dib.pkl and scaler_hist_dib.pkl) using joblib, preparing them for deployment in an application like Streamlit.

6. 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-Scor	TN	FP	FN	TP
Stacked (XGB → LR)	0.852	0.317	20,322	2,209	1,474	353
Logistic Regression	0.857	0.318	20,476	2,055	1,499	328
Decision Tree	0.850	0.227	20,591	1,940	1,782	545

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

i) Captures complex feature interactions, non-linear relationships, and threshold effects.

ii) Generalize better on new unseen data.
iii) Avoid underfitting (especially if relationships are nonlinear or interaction-based).
iv) Detects 25 more true disease cases, which is crucial for *medical or chronic disease prediction*

Model 2: Diabetes Detection result

Model	Accuracy	F1-Score (Weight 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

7. Challenges and Future Improvements

Challenges:

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

Future Enhancements

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

8. 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.

Chronic Disease Screening — WebApp

Overview This Streamlit app screens for chronic conditions using pre-trained scikit-learn models, caching loaded artifacts for fast inference and rendering probability charts with Altair. Artifacts are loaded via joblib and should be version-pinned because scikit-learn pickles are not guaranteed to be cross-version compatible.

Features

- Sidebar form collects demographics and lifestyle inputs using a submit button so values update in a single batch.
- BMI is computed from height and weight and shown alongside the standard adult BMI categories for context.
- Diabetes head: multiclass probabilities are displayed for Diabetic and Prediabetic with the top class in the header and a 0–100% bar scale for readability.
- Heart head: binary prediction shows the chosen class and the positive-class probability with a single bar on a 0–100% scale.
- Models and scalers are cached with st.cache_resource to avoid reloading on each rerun and to share resources across sessions.

Requirements Python environment with Streamlit, scikit-learn, joblib, pandas, numpy, and Altair installed and version-pinned alongside your pickled models.

9. References

- Centers for Disease Control and Prevention. (2025, August). 2024 BRFSS survey data and documentation. CDC. https://www.cdc.gov/brfss/annual_data/annual_2024.html
- OpenAI. (2025). *ChatGPT* (GPT-5) [Large language model]. https://chat.openai.com/
- Google. (2025). Colab AI [Computer software]. Google. https://colab.research.google.com/