Project Report: Diabetes Health Indicators Analysis

1. Introduction

Diabetes is one of the most common chronic diseases worldwide and poses a significant public health concern. Early detection and prevention are crucial because undiagnosed or poorly managed diabetes can lead to severe health complications, including cardiovascular disease, kidney failure, blindness, and nerve damage.

Accurate prediction of diabetes risk allows healthcare providers to intervene earlier, promote lifestyle changes, and improve disease management outcomes. With the availability of large-scale health surveillance data, machine learning models can now be used to identify potential predictors of diabetes and their interactions with lifestyle and demographic factors.

---

2. Understanding Diabetes

There are primarily two major types of diabetes:

- **Type 1 Diabetes:** An autoimmune condition in which the immune system attacks insulin-producing cells in the pancreas. It typically occurs early in life and requires lifelong insulin therapy.

- **Type 2 Diabetes:** A metabolic disorder where the body becomes resistant to insulin or fails to produce enough insulin. It is strongly associated with obesity, inactivity, and poor dietary habits and is more common in adults.

If diabetes goes undiagnosed or unmanaged, it can lead to long-term complications affecting nearly every organ in the body. Hence, identifying risk factors and promoting early detection are essential for public health.

---

3. Data Source

The dataset used for this project comes from the **Centers for Disease Control and Prevention (CDC)**, collected through the **Behavioral Risk Factor Surveillance System (BRFSS)**.
The BRFSS is the nation's premier system of health-related telephone surveys that collect data

from U.S. residents about their health-related risk behaviors, chronic health conditions, and preventive service use.

---

4. Data Format and Conversion

The BRFSS data is originally provided in the **.XPT (SAS Transport) format**, which is a standardized file type developed by SAS to store and exchange datasets between different software environments.
To make the data accessible for analysis using Python and modern data science tools, we converted it into a **.CSV (Comma-Separated Values)** format using a custom Python script.

---

5. Dataset Overview

After conversion, the dataset contained:

- **Rows:** 457,670
- **Columns:** 301

The dataset is extensive, covering numerous aspects of health, lifestyle, and demographics. Each record corresponds to a respondent from the BRFSS telephone survey.

---

6. Data Structure

The dataset includes variables from multiple sections of the BRFSS survey, such as:

- Adverse Childhood Experiences
- Alcohol Consumption
- Arthritis
- Cancer Screening and Survivorship
- Chronic Health Conditions
- Cognitive Decline
- Demographics
- Diabetes
- Exercise and Physical Activity
- Family Planning
- Health Care Access

- Health Status
- HIV/AIDS and HPV Vaccination
- Oral Health
- Tobacco Use
- Prostate and Colorectal Cancer Screening
- Immunization and more

This wide coverage makes the dataset suitable for exploring diverse health and behavioral patterns linked to diabetes risk.

---

7. Data Cleaning and Preprocessing

The initial dataset contained a large number of irrelevant or incomplete columns. The following steps were taken to prepare it for analysis:

1. **Removal of Unnecessary Columns:**
   We eliminated metadata fields such as date fields, version identifiers, and state-specific question columns that were not directly relevant to diabetes prediction.

2. **Handling Missing Data:**
   Many columns contained missing or incomplete responses, as not all questions were asked across all states. We removed columns with excessive missing values (ie >40%) and replaced certain coded values with `NaN` for clarity.

3. **Dealing with High-Null Columns:**
   For example, the *Adverse Childhood Experiences (ACE)* section had approximately **91% missing values**. While this limited its use in model training, we still conducted exploratory analysis to understand potential relationships.

4. **Data Consistency and Aggregation:**
   Related features (e.g., smoking habits, asthma-related questions, income levels,other chronic diseases) were aggregated into single representative columns to simplify the dataset.

After this stage, we reduced the dataset from 301 columns to **92 significant variables**, and after further filtering, we finalized **51 key features** for further feature selection using modeling.

---

8. Feature Selection and Statistical Analysis

To identify which variables most strongly influenced the target variable **DIABETE4** (the diabetes status), we applied statistical tests:

- **For Categorical vs. Categorical Variables:**
  We used **Chi-Square tests**, along with **Cramer's V** and **p-values**, to measure association strength.

- **For Categorical vs. Numerical Variables:**
  We applied **Spearman's rank correlation** and other appropriate tests to detect monotonic relationships.

During the data analysis process, we observed that not all states administered the same set of questions in the BRFSS survey. Each state selected specific optional modules to include, meaning certain questions were only asked in a subset of states. As a result, several features contained very high proportions of missing values because they were not part of every state's questionnaire.

For example, modules such as **Adverse Childhood Experiences (ACE)** and **Marijuana Use** were only asked in a limited number of states. Although these features were potentially important for understanding the broader determinants of diabetes, the large number of missing responses (often exceeding 85–90%) would have introduced significant bias and data imbalance into the model.

Therefore, we made a decision to **exclude these variables** from the predictive modeling process. This ensured that the dataset remained statistically reliable and representative of all respondents, rather than being skewed toward the few states that included these modules.

---

9. Final Dataset

After systematic filtering and feature engineering, the final dataset used for model training contained:

- **Rows:** Subset of cleaned, valid responses
- **Features:** 51 well-defined, meaningful features
- **Target Variable:** DIABETE4 - HasDiabetes (binary classification indicating diabetes status)

10. Modeling Approach

Objective

To make the model for our project we have to make a predictive model that estimates the likelihood of diabetes based on demographic, lifestyle, and health indicators from the BRFSS dataset.

Approach Summary

- We used supervised binary classification as provided in the dataset.
- We applied preprocessing pipelines for scaling numeric and encoding categorical features.
- We decided to handle class imbalance through stratified sampling and class-weight adjustments in the dataset.

  Compared multiple models: we compare the model provided below with our cleaned database for evaluation.

- Logistic Regression – Interpretable baseline .
- Random Forest – Ensemble learning for non-linear patterns.
- CatBoost – Handled categorical variables and class imbalance efficiently.

11. Model Training and Evaluation

Feature Selection Method:

Embedded method using CatBoost model-based feature importance.
Importance is derived from each feature's contribution to reducing model loss during tree splits.
The top 20 features were retained to simplify the model and focus on the most predictive variables.

Hyperparameter Tuning:

Parameters such as learning rate, tree depth, regularization strength, and iteration count were tuned manually for stable convergence and generalization.
Early stopping (200 rounds) was used as an automatic regularization mechanism to prevent overfitting.

Train-Test Split

Parameters such as learning rate, tree depth, regularization strength, and iteration count were tuned manually for stable convergence and generalization.
Early stopping (200 rounds) was used as an automatic regularization mechanism to prevent overfitting.

We divided the data into two parts. 80% of the dataset was set for training the model and 20% for testing it. We made sure the split kept the same ratio of diabetic and non-diabetic people in both parts, so the model learns fairly.

In our dataset, about 87% of people do not have diabetes and about 13% do. We kept this balance in both the training and testing sets.

Evaluation Metrics

- Accuracy, F1-Score, and ROC-AUC used for performance comparison.
- Confusion Matrix and Classification Report provided detailed insights.

Model	**Accuracy**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.73	0.43	0.82
Random Forest	0.86	0.11	0.81
CatBoost	0.84	0.47	0.88

12. Feature Importance and Medical Validation

Key Findings

Top predictors from permutation importance and correlation analysis:

1. **Age Group**
2. **BMI**
3. **Physical Health Status**
4. **Mental Health Status**
5. **Smoking Habits**
6. **Drinking Habits**
7. **Race**
8. **Chronic Diseases**
9. **General Health status**

Medical Validation

- High BMI and poor physical health are known clinical risk factors for diabetes.

- Smoking, drinking and age are well-established lifestyle contributors.
- Mental health correlations suggest psychosocial impacts on chronic illness risk.

The model aligns closely with real-world medical evidence, increasing interpretability and trustworthiness.

13. Future Scope and Learnings

Through this project, we identified several important lessons and future directions that could improve the reliability and interpretability of diabetes prediction models.

Learnings:

1. Need for More Comprehensive and Consistent Data

The dataset used (BRFSS) is large and diverse but not uniform across all states. Many questions were only asked in selected states, resulting in **high proportions of missing values** for certain variables.
Future data collection should aim for **more comprehensive coverage**, ensuring that all participants are asked the same set of core health and lifestyle questions.
This consistency would enable models to use a richer and more complete set of predictors, leading to better accuracy and generalizability.

2. Limitation: Data Represents People Already Diagnosed

An important observation is that most participants who reported diabetes in the BRFSS survey were individuals **already aware of their diagnosis**.
Such respondents often **modify their lifestyle and health behaviors** after diagnosis — for example, improving diet, increasing physical activity, or quitting smoking.
As a result, their current health and lifestyle variables may not accurately represent the conditions **that led to diabetes**, but rather reflect their **post-diagnosis management behaviors**.

This introduces a challenge: our model might be learning patterns that describe ***how people live with diabetes*** **rather than** ***how diabetes develops*****.**

Future studies should therefore include **longitudinal or pre-diagnosis data** to better capture the true risk factors that precede diabetes onset.

3. Unclear Causality Between Diseases and Diabetes

While several chronic diseases and health conditions (e.g., hypertension, obesity, heart disease) show strong correlations with diabetes, **correlation does not imply causation**.
It is often unclear whether these conditions **contributed to the onset of diabetes** or **resulted from it**.

Without knowing the sequence of events or time of diagnosis, it becomes difficult to determine causality.

Future datasets should include **temporal or event-based data**, for example, when each condition was first diagnosed — so that causal relationships can be explored more effectively.

4. Ideal Future Dataset

For truly predictive and actionable modeling, we would need data that:

- Tracks **individuals over time** (before and after diabetes diagnosis).
- Captures **lifestyle, habits, and biological measures** *before* they were aware of their condition.
- Minimizes missing values by ensuring uniform question sets across all states and modules.
- Includes **timestamped or sequential health information** to study cause–effect dynamics.

Such data would allow us to build models capable not **just of identifying diabetes**, but of predicting **who is likely to develop it** — which is the real goal of preventive healthcare analytics.

Future Enhancements:

- **Collect Longitudinal and Pre-Diagnosis Data:**
  Obtain data tracking individuals **before, during, and after** diabetes diagnosis to identify genuine causal factors rather than post-diagnosis behaviors.
- **Enhance Data Consistency Across States:**
  Advocate for **uniform survey modules** across all states to reduce missing data and improve feature coverage.
- **Expand to Predict Pre-Diabetes and Disease Progression:**
  Extend the model to detect **early risk or pre-diabetic states** and forecast future disease development.
- **Develop a Public Health Awareness Dashboard:**
  Create an interactive dashboard that visualizes key risk factors, state-wise diabetes trends, and lifestyle correlations for **public education and policy insights**.
- **Periodically Retrain Model with Updated BRFSS Data:**
  Continuously update and retrain the model with new annual BRFSS data to maintain relevance and adapt to **changing population behaviors and health pattern**