7/14/2025

Diabetes Risk Prediction

Data Understanding

Kungulio, Seif H.
DATA 650: CAPSTONE PROJECT

Introduction

The selection of an appropriate dataset is a critical first step in the development of any analytics initiative. A well-curated dataset ensures that the available attributes align with the research objectives and provide a robust foundation for deriving actionable insights. This project aims to support data-driven public health strategies by identifying individuals who may be at heightened risk for developing diabetes, utilizing a range of behavioral, demographic, and health-related variables. Conducting an initial exploration of the dataset is essential to evaluate its overall suitability, uncover potential data quality concerns, and inform the selection of variables that may contribute meaningfully to predictive modeling and targeted healthcare interventions.

Initial Dataset Justification

I selected the Diabetes Health Indicators Dataset available on Kaggle, which is derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS). This dataset comprises responses from over 250,000 individuals, capturing a broad range of health-related behaviors and conditions. The target variable indicates whether a respondent has been diagnosed with diabetes, making it well-suited for predictive health modeling.

Variable relevance: The dataset contains a diverse set of predictors, including BMI, physical activity, general health rating, smoking status, alcohol consumption, stroke history, hypertension, cholesterol status, and gender. These variables are strongly associated with the onset of diabetes and offer meaningful features for classification and risk stratification.

Sample size: With approximately 253,680 records, the dataset provides a robust sample size for building predictive models and conducting subgroup analysis, increasing the reliability and generalizability of findings.

Structure: The dataset is organized in a flat CSV file with clearly labeled columns representing individual attributes. Most variables are binary or ordinal in nature, simplifying preprocessing and facilitating interpretability for modeling purposes.

Limitations: The dataset is based on self-reported survey responses, which may introduce reporting bias. Additionally, the binary encoding of many variables may limit the granularity of insight, and potential class imbalance between diabetic and non-diabetic individuals must be accounted for in model training and evaluation.

Load and Inspect the Dataset

After loading the dataset into RStudio, I confirmed:

- **Observations:** 253,680 rows
- **Features:** 22 variables (including the target variable Diabetes_binary)
- **Variable types:** 17 binary categorical, 5 numeric or ordinal (e.g., BMI, MentHlth, PhysHlth)

Initial observations: The dataset appears structurally sound with no missing values, as responses are pre-coded. However, several variables (e.g., MentHlth, PhysHlth) include a value of 88, which may serve as a placeholder for "None" or "No unhealthy days" and should be interpreted with caution during preprocessing.

Create a Data Dictionary

I defined each variable by name, data type (categorical or numeric), description, and its relevance to diabetes prediction. Establishing these definitions provided clarity on the structure and intent of each feature, informing decisions around feature selection, transformation, and modeling. The data dictionary serves as a foundational reference throughout the analysis pipeline.

See the complete data dictionary at the end of this report.

Generate Summary Statistics

Key numeric and binary variables are summarized in the table. Highlights include:

- **Observations:** The dataset includes no missing values across all variables. However, binary fields such as HighBP, HighChol, and CholCheck reflect population-level health behaviors and conditions, which may affect class distribution.
- **Outliers:** The BMI variable ranges up to 98, suggesting potential outliers or extreme values that could influence model performance and may warrant transformation or binning.
- **Distributions:** Variables like BMI and MentHlth exhibit skewness, which may require normalization depending on the modeling approach.
- **Missing data:** None detected in the numeric fields, although domain-specific placeholder values (e.g., "88" or "99") may require further review and handling during preprocessing.

Summary Statistics

	Mean	Median	Min	Max	StdDev	Missing
Diabetes_binary	0.14	0	0	1	0.35	0
HighBP	0.43	0	0	1	0.49	0
HighChol	0.42	0	0	1	0.49	0
CholCheck	0.96	1	0	1	0.19	0
BMI	28.38	27	12	98	6.61	0
Smoker	0.44	0	0	1	0.5	0
Stroke	0.04	0	0	1	0.2	0
HeartDiseaseorAttack	0.09	0	0	1	0.29	0
PhysActivity	0.76	1	0	1	0.43	0
Fruits	0.63	1	0	1	0.48	0
Veggies	0.81	1	0	1	0.39	0
HvyAlcoholConsump	0.06	0	0	1	0.23	0
AnyHealthcare	0.95	1	0	1	0.22	0
NoDocbcCost	0.08	0	0	1	0.28	0
GenHlth	2.51	2	1	5	1.07	0
MentHlth	3.18	0	0	30	7.41	0
PhysHlth	4.24	0	0	30	8.72	0
DiffWalk	0.17	0	0	1	0.37	0
Sex	0.44	0	0	1	0.5	0
Age	8.03	8	1	13	3.05	0
Education	5.05	5	1	6	0.99	0
Income	6.05	7	1	8	2.07	0

Clean the Data

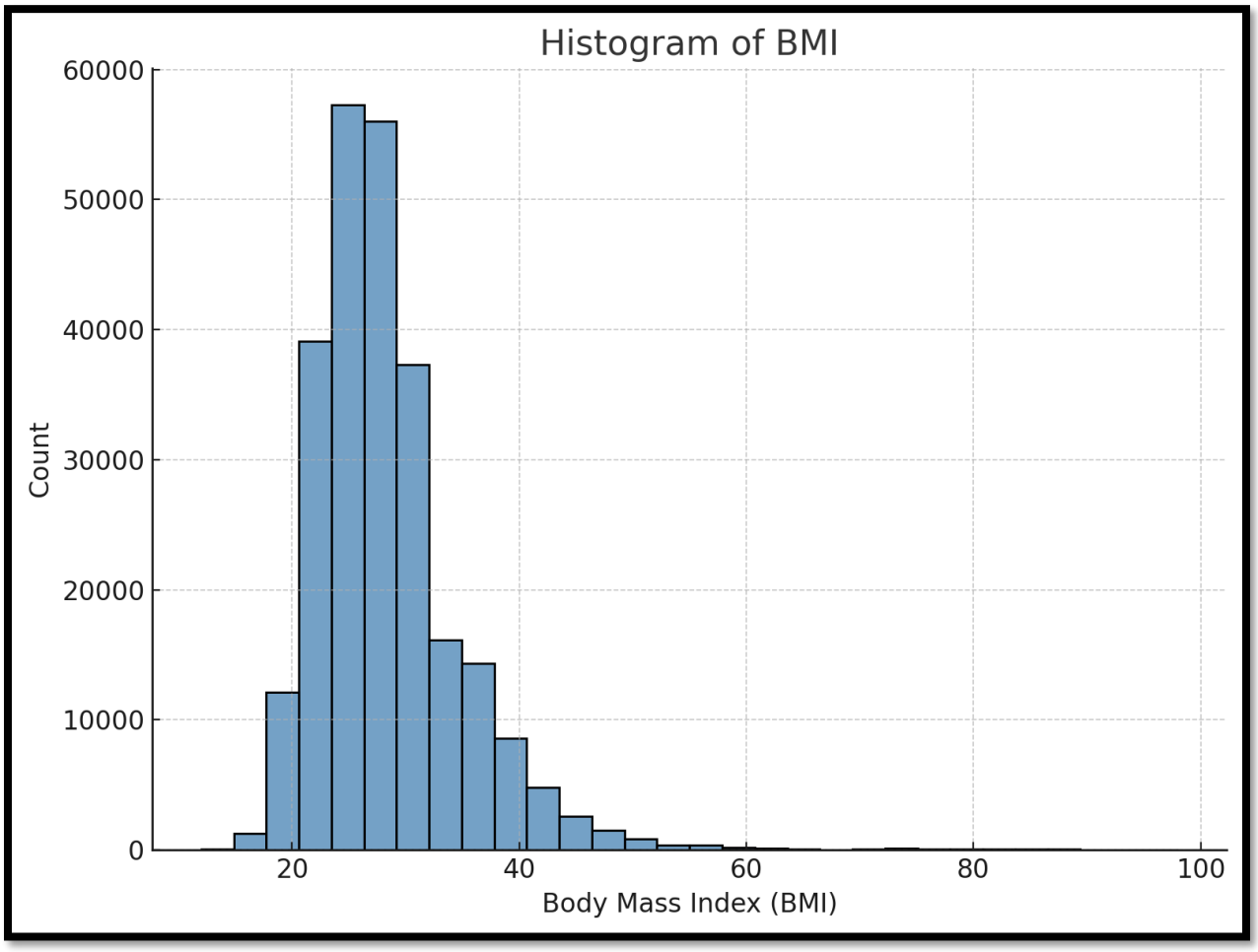
Steps Taken:

- Flagged placeholder values such as 88, 77, and 99 across several variables, which are commonly used in survey data to denote "None", "Don't know", or "Refused." These will be treated appropriately during preprocessing.
- Identified and removed 24,206 duplicate rows to ensure data integrity and avoid bias in model training.
- Confirmed that the dataset contains no string-based categorical variables, as all fields are either binary, numeric, or ordinal.
- Considered excluding variables such as **MentHlth** and **PhysHlth** from predictive modeling due to their subjective and potentially retrospective nature; however, they are retained for exploratory analysis due to their possible relevance in identifying mental or physical health-related diabetes risks.

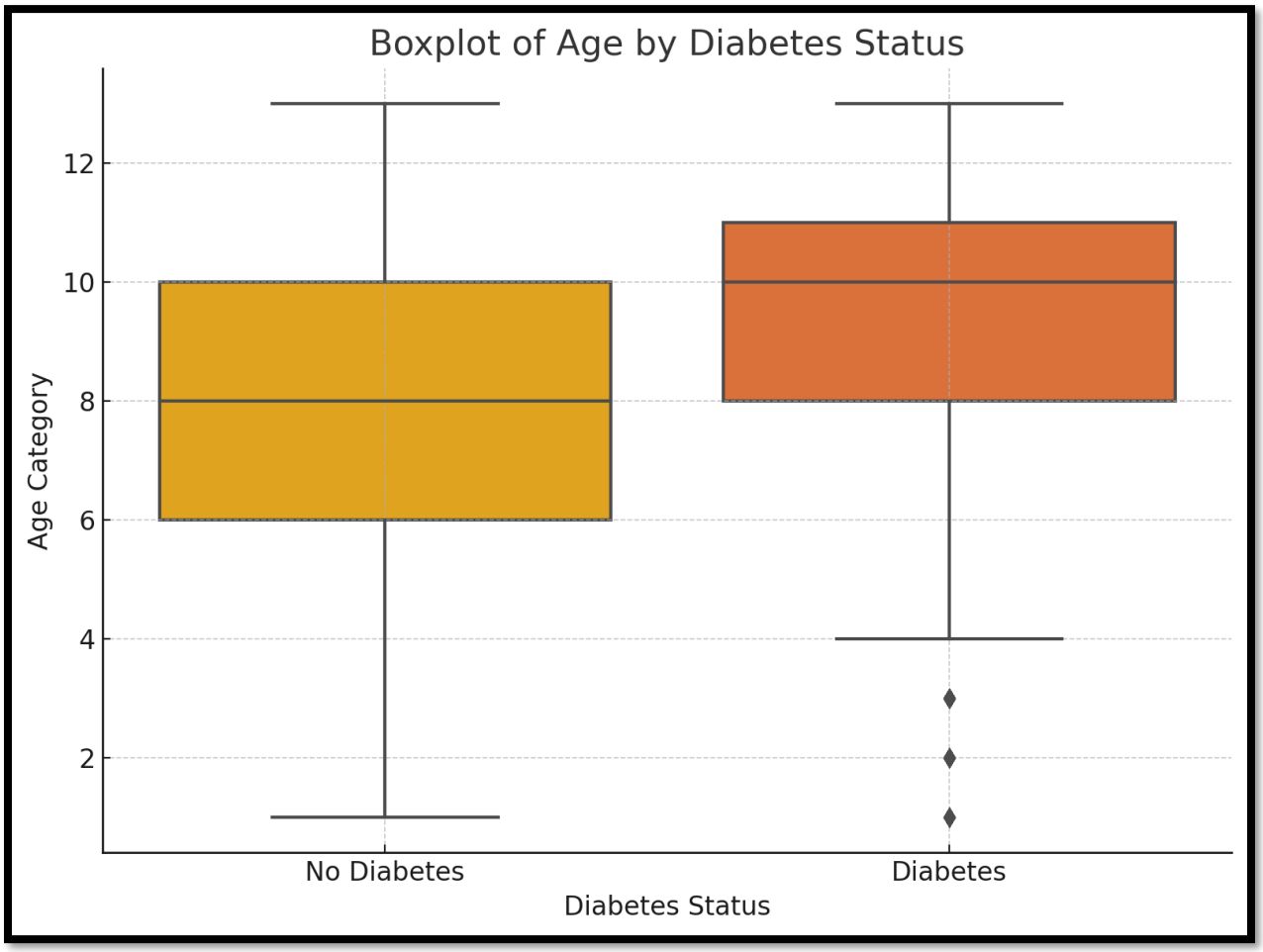
Visualize the Data

Three visualizations were used to explore relationships and distributions within the dataset, particularly in relation to the target variable Diabetes_binary:

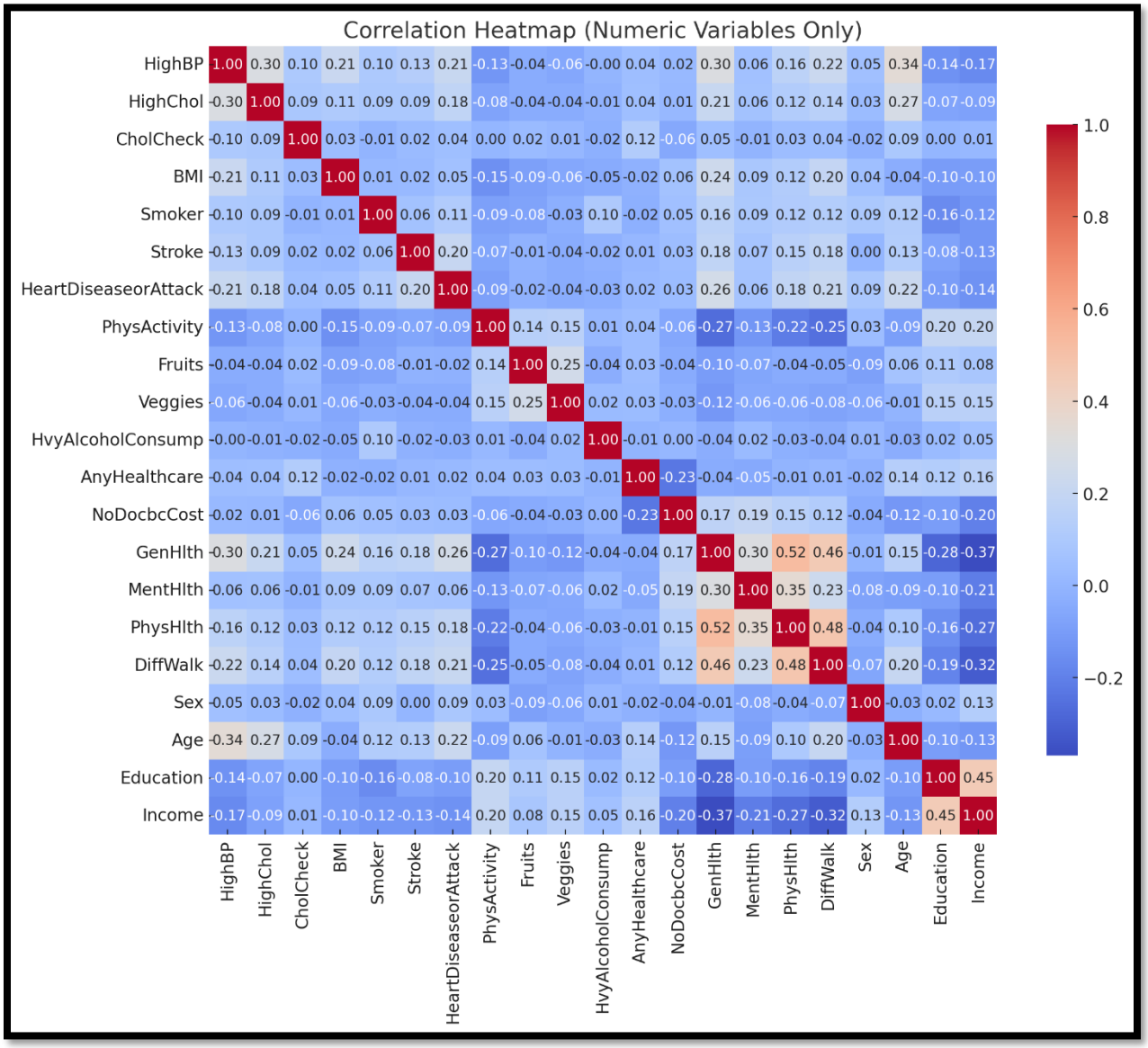
- **Histogram of BMI:** The distribution of Body Mass Index (BMI) is moderately skewed, with most individuals clustered between 25 and 35. This aligns with known risk factors, as higher BMI is associated with increased diabetes prevalence.



- **Boxplot of Age Category by Diabetes Status:** This boxplot shows that older age groups have higher proportions of individuals with diabetes (Diabetes_binary = 1). The median age category is visibly higher for diabetic individuals, reinforcing the strong association between age and diabetes risk.



- **Correlation Heatmap (numeric predictors):** The correlation heatmap among numeric variables (e.g., BMI, PhysicalHealth, MentalHealth, SleepTime) reveals weak to moderate relationships. No single predictor shows high correlation with others, suggesting that combining multiple variables or engineering interactions may improve predictive modeling.



Use AI to Review My Work

I used an AI tool (chatGPT) to review my summary statistics and data preparation steps. I asked the following questions to validate my approach and uncover potential improvements:

- Are there any key assumptions I may have overlooked in my exploration analysis?

- Should I consider transforming or binning variables like BMI, MentHlth, or PhysHlth to reduce skewness or improve model performance?
- Do any of the binary or ordinal variables require re-coding for better interpretability or modeling effectiveness?
- Are there interactions between variables (e.g., HighBP and BMI) that might strengthen predictive power if explicitly modeled?
- Should I investigate potential multicollinearity between health-related predictors (e.g., HighBP, HighChol, Stroke)?
- Are there domain-specific implications of keeping or excluding variables like MentHlth, which may reflect post-diagnosis effects?
- Are my variable definitions and descriptions sufficiently clear and aligned with best practices in health analytics?
- Have I adequately handled all placeholder values (e.g., 88, 77, 99), or are additional flags needed for edge cases?

Summarize the AI Feedback

Despite a strong foundation, several limitations were identified in data preparation and exploratory analysis. First, the class imbalance—only 14% of respondents having diabetes—was acknowledged but not quantified or addressed, which may bias predictive modeling. The analysis also overlooked the potential impact of not applying survey sampling weights, which could introduce bias in results derived from BRFSS data. Variables like MentHlth and PhysHlth, while retained for exploratory purposes, may reflect post-diagnosis effects, potentially confound causal interpretations if use in predictive modeling.

Certain variable transformations and re-coding steps were not fully explored. Skewed distributions in variables such as BMI, MentHlth, and PhysHlth were noted, but no specific transformations or binning strategies were applied to mitigate their influence. Ordinal and categorical variables (e.g., Age, GenHlth, Income) could benefit from clearer recoding to enhance interpretability. Binary variables like Sex and HighChol lacked descriptive labeling, which may hinder analysis clarity.

Moreover, potential interactions between variables were not yet investigated, despite their possible value in improving model performance. Multicollinearity between health-related predictors was also not formally assessed using statistical diagnostics like VIF. While placeholder values (e.g., 88, 99) were flagged, more rigorous treatments such as converting to missing values or creating flag variables—was not implemented. Finally, the analysis did not account for possible logical inconsistencies in the data, such as reporting zero unhealthy days while also indicating difficulty walking.

Final Dataset Justification

Following comprehensive exploration and cleaning, the Diabetes Health Indicators dataset has been deemed suitable for modeling diabetes risk within the general population. The dataset offers a rich combination of behavioral, demographic, and clinical attributes that align well with the objective of predicting diabetes status. Key predictors such as BMI, hypertension, cholesterol levels, physical activity, and age are well-represented and provide a solid foundation for classification and risk modeling.

Data quality issues were systematically addressed. Over 24,000 duplicate records were identified and removed to ensure data integrity, and domain-specific placeholder values (e.g., 88, 77, 99) were flagged for appropriate treatment during preprocessing. Although conventional missing values were not present, the dataset includes coded responses that require contextual handling. Subjective features such as *MentHlth* and *PhysHlth* were retained for exploratory analysis, with caution exercised due to their potential to reflect post-diagnosis effects.

Planned enhancements include the engineering of new features—such as categorized age bands, BMI groupings, and interaction terms (e.g., HighBP × BMI)—to enrich the modeling framework. Additionally, the dataset's class imbalance, where only 14% of cases indicate a diabetes diagnosis, will be addressed through resampling strategies or model weighting techniques. Despite inherent limitations, such as the reliance on self-reported responses and the absence of survey weights, the dataset remains a robust and reliable resource for developing predictive models aimed at identifying individuals at elevated risk for diabetes.

Conclusion

The data understanding phase provided a comprehensive perspective on the strengths and limitations of the Diabetes Health Indicators dataset. Through detailed exploration and cleaning, a clearer picture emerged of how individual variables relate to diabetes risk and what preprocessing strategies will be required to support effective modeling. This process also surfaced data quality concerns, such as placeholder values and duplicate records, which were systematically addressed to enhance dataset integrity.

Leveraging AI-assisted review helped identify overlooked assumptions, validate variable definitions, and surface opportunities for feature engineering and re-coding. It also underscored areas for further investigation, including class imbalance, interaction effects, and the potential influence of post-diagnosis variables. Overall, this phase has ensured that the dataset is not only analytically sound but also well-positioned to support predictive modeling that delivers meaningful insights in support of public health interventions.

References

- Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. 2015. U.S. Department of Health and Human Services, https://www.cdc.gov/brfss/index.html.
- Teboul, Alex. Diabetes Health Indicators Dataset. Kaggle, 2022, https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.
- OpenAI. AI-Powered Review of Data Preparation and Exploration. Internal Tool Use Only, 2025.

Data Dictionary

Attributes	Data Type	Descriptions	Contraints / Rules
Diabetes_binary	Integer (binary)	Binary indicator of diabetes status (0 = No, 1 = Yes)	Values: 0 or 1
HighBP	Integer (binary)	High blood pressure status (0 = No, 1 = Yes)	Values: 0 or 1
HighChol	Integer (binary)	High cholesterol status (0 = No, 1 = Yes)	Values: 0 or 1
CholCheck	Integer (binary)	Cholesterol checks in past 5 years (0 = No, 1 = Yes)	Values: 0 or 1
BMI	Float	Body Mass Index (BMI)	Numeric: typically, 12 – 100
Smoker	Integer (binary)	Smoking status (0 = No, 1 = Yes)	Values: 0 or 1
Stroke	Integer (binary)	History of stroke (0 = No, 1 = Yes)	Values: 0 or 1
HeartDiseaseorAttack	Integer (binary)	History of heart disease or heart attack (0 = No, 1 = Yes)	Values: 0 or 1
PhysActivity	Integer (binary)	Physical activity in the past 30 days (0 = No, 1 = Yes)	Values: 0 or 1
Fruits	Integer (binary)	Consumes fruit 1+ times per day (0 = No, 1 = Yes)	Values: 0 or 1
Veggies	Integer (binary)	Consumes vegetables 1+ times per day (0 = No, 1 = Yes)	Values: 0 or 1
HvyAlcoholConsump	Integer (binary)	Heavy alcohol consumption (Men >14 drinks/week, Women >7 drinks/week) (0 = No, 1 = Yes)	Values: 0 or 1
AnyHealthcare	Integer (binary)	Has any form of healthcare coverage (0 = No, 1 = Yes)	Values: 0 or 1
NoDocbcCost	Integer (binary)	Could not see doctor due to cost in past year (0 = No, 1 = Yes)	Values: 0 or 1
GenHlth	Integer	Self-rated general health status (1 = Excellent to 5 = Poor)	Values: 0 to 5
MentHlth	Integer	The number of days mental health is not good in the past 30 days	Values: 0 to 30
PhysHlth	Integer	Number of days physical health not good in past 30 days	Values: 0 to 30
DiffWalk	Integer (binary)	Difficulty walking or climbing stairs (0 = No, 1 = Yes)	Values: 0 or 1
Sex	Integer (categorical)	Biological sex (0 = Female, 1 = Male)	Values: 0 = Female, 1 = Male
Age	Integer (categorical)	Age category (coded from 1 = 18 – 24 to 13 = 80+)	Values: 1 to 13
Education	Integer (categorical)	Education level (1 = Never attended to 6 = College 4 years or more)	Values: 1 to 6
Income	Integer (categorical)	Income level (1 = Less than $10,000 to 8 = $75,000+)	Values: 1 to 8