7/14/2025

**Kungulio, Seif H.**

**DATA 650: CAPSTONE PROJECT**

**Diabetes Risk Prediction**

**Data Understanding**

## **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](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download), 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 77, 88, 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**

## **Use AI to Review My Work**

## **Summarize the AI Feedback**

## **Final Dataset Justification**

## **Conclusion**