Building Risk Prediction Models for Diabetes Using Machine Learning

Phase 1 Project Update

Presented by:
Leslie Li, Satyaki Dixit, and Shreshta Phogat

BACKGROUND

Severity

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy.

Scale

The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have pre-diabetes. Diagnosed diabetes cost roughly $327 million dollars and total costs with undiagnosed diabetes and pre-diabetes approaching $400 billion dollars annually.

Significance

Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

LITERATURE OVERVIEW

- University of Rochester School of Medicine and Dentistry built risk prediction models for Type 2 Diabetes using supervised ML models such as SVM, Decision Tree, and Logistic Regression models. (Xie et al, 2019)
- Department of Mathematics and Statistics from York University used threshold method and the class weight to improve sensitivity - the proportion of diabetes patients correctly predicted by models such as Decision Tree and Random Forest. (Lai et al, 2019)
- Department of Endocrinology and Metabolism from Peking University People's Hospital found that sex, age, history of diabetes, waist circumference, BMI, SBP were important risk factors related to diabetes. (Zhou et al, 2013)
- Insufficient sleep duration and/or sleep restriction in the laboratory, poor sleep quality, and sleep disorders such as insomnia and sleep apnea have all been associated with diabetes risk (Grandner, 2016).

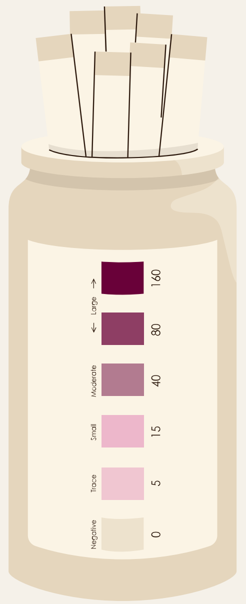
CHALLENGE 1

There is considerable heterogeneity in previous studies regarding machine learning techniques used, making it challenging to identify the optimal one.

CHALLENGE 2

There is a lack of transparency about the features used to train the models, which reduces their interpretability, a feature utterly relevant to the doctor.

INTRODUCTION

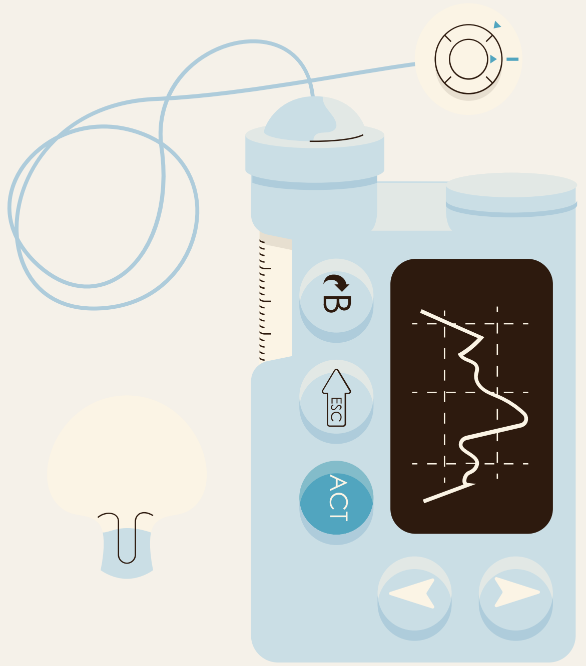


Data Source

- The Behavioral Risk Factor Surveillance System's survey responses in 2015.
- Health-related telephone surveys collecting state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

Research Question

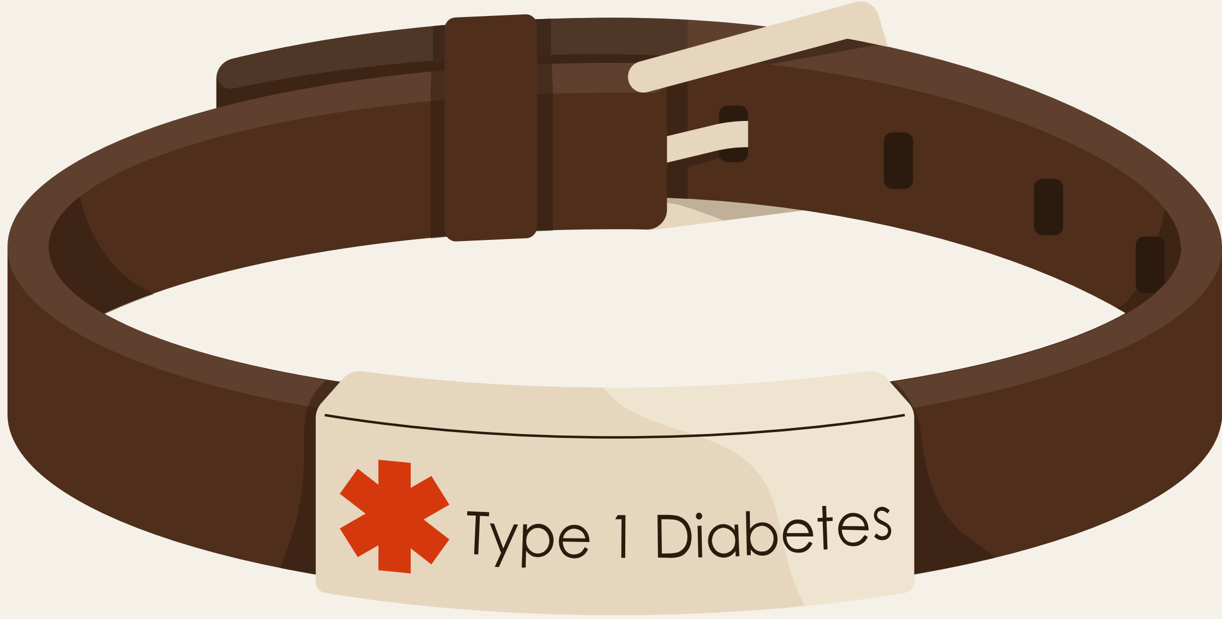
- What risk factors are most predictive of diabetes risk?
- What is the association among different variables?
- Which ML models contribute to more accurate prediction?
- What are the optimal validation metrics to measure model performance?



Methodology

- Select essential risk factors for analysis after literature review
- EDA with dichotomy and transformation
- Use multivariable weighted logistic regression models to measure associations among factors
- Apply supervised ML models and metrics

DATA OVERVIEW



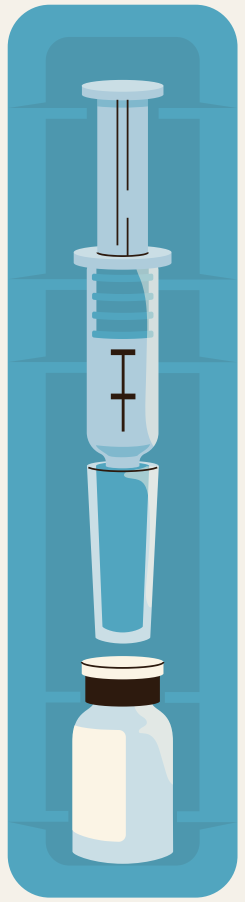
Shape

- 330 features (columns)
  - 323 numerical features
  - 7 categorical features
  - 244 columns have missing values
- 441,456 survey responses (rows)
- Not balanced with a size at 541.28 MB

A Glimpse of Attributes

- High BP
- High cholesterol, cholesterol check
- BMI
- Smoke history, stoke history
- Coronary heart disease (CHD) or myocardial infarction
- Physical activity in past 30 days
- Fruit, vegetables, drinks consumption habit
- Health care coverage, doctor visit frequency, health scale
- Mental health
- Sex, age, education, income level
- Sleep/disordered breathing

PLAN OF ACTION



Step 1

Clean BRFSS data into a useable format, including all important features relating to diabetes risk for machine learning algorithms

Started

Step 2

Exploratory data analysis/ feature selection: Determining the features that have the most impact on diabetes onset/risk

Ongoing

Step 3

Implement machine learning models using a subset of risk factors to predict diabetes risk

Planned

DATA CLEANING

Starting point:

- 2015 data with 330 Columns and ~400K+ rows

Manipulations:

- Chose 41 most relevant columns based on known relation to diabetes risk (Mostly calculated variables to simplify the process)
- Clean each variable down to a simplified scale using CDC provided CodeBook report
- Make the feature names more readable

DATA CLEANING EXAMPLE

BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM
CODEBOOK REPORT, 2015
Land-Line and Cell-Phone data

Respondents aged 18-64 with health care coverage

CalculatedVariables: 3.1 Calculated Variables **Type:** Num

Column: 1895 **SAS Variable Name:** _HCVU651

Prologue:

Description: Respondents aged 18-64 who have any form of health care coverage

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Have health care coverage Notes: 18 <= AGE <=64 and HLTHPLN1 = 1	252,542	57.21	67.27
2	Do not have health care coverage Notes: 18 <= AGE <=64 and HLTHPLN1 = 2	29,661	6.72	11.67
9	Don't know/Not Sure, Refused or Missing Notes: AGE > 64 or AGE = Missing or HLTHPLN1 = 7 or 9 or Missing	159,253	36.07	21.05

High Blood Pressure Calculated Variable

CalculatedVariables: 4.1 Calculated Variables **Type:** Num

Column: 1896 **SAS Variable Name:** _RFHYPE5

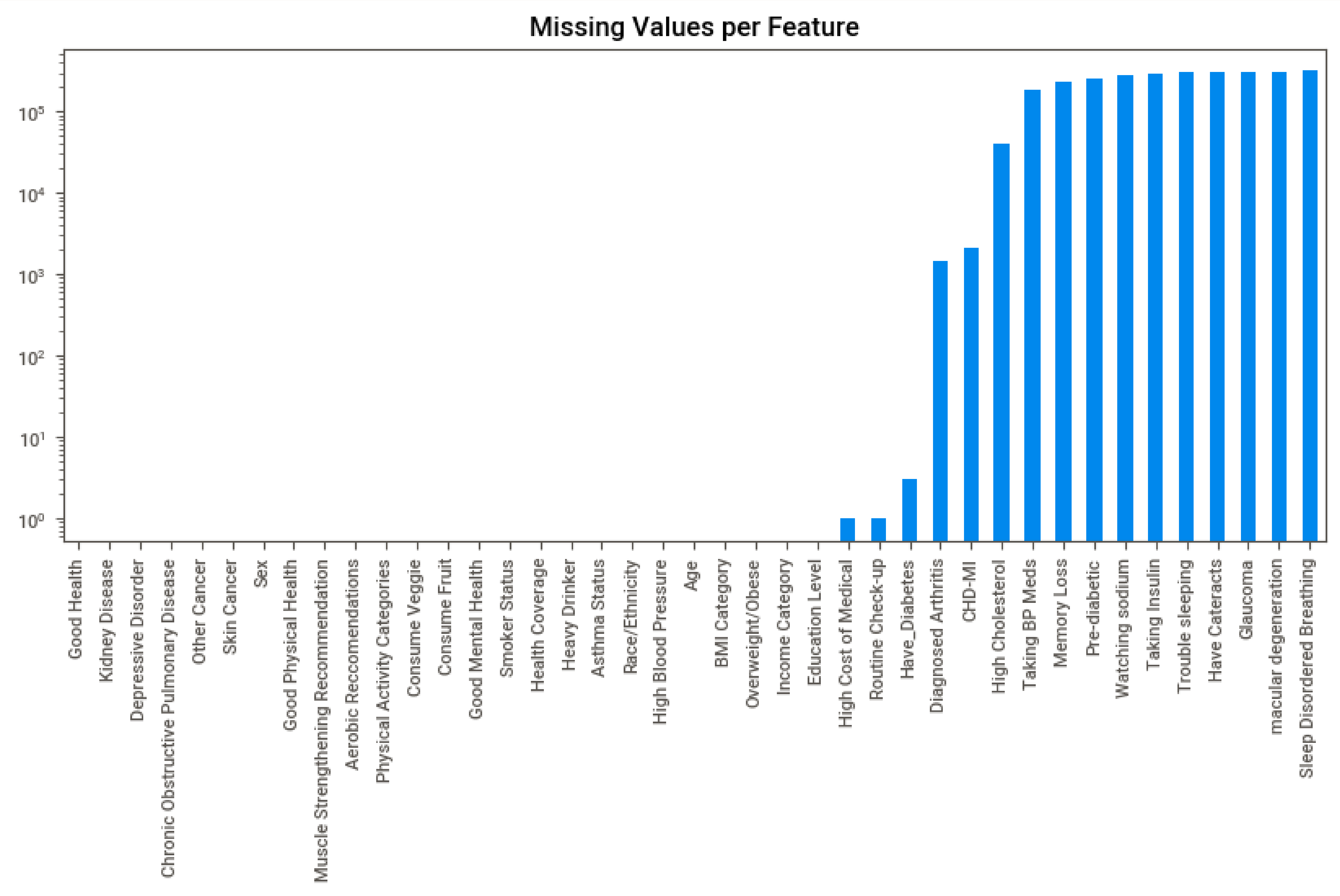
Prologue:

Description: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No Notes: BPHIGH4 = 2 or 3 or 4	261,901	59.33	67.78
2	Yes Notes: BPHIGH4 = 1	178,188	40.36	31.90
9	Don't know/Not Sure/Refused/Missing Notes: BPHIGH4 = 7 or 9 or Missing	1,367	0.31	0.31

- Questions on the health survey were condensed into calculated variables and given arbitrary values
- For example, health coverage and high blood pressure for both of these features 1 and 2 mean completely different things
- We changed having health coverage and having high blood pressure to 1 and not having to mean 2
- This analysis will be done for all 41 variables
- Then all feature names will be made readable

EDA - QUALITY INVESTIGATION



- Scale unique value count for each feature logarithmically
- 13 columns have missing data and great values concentrate on features:
    - Pre-diabetic
    - Taking Insulin
    - Macular Degeneration
    - Memory Loss
    - Sleep Disordered Breathing
    - Trouble Sleeping

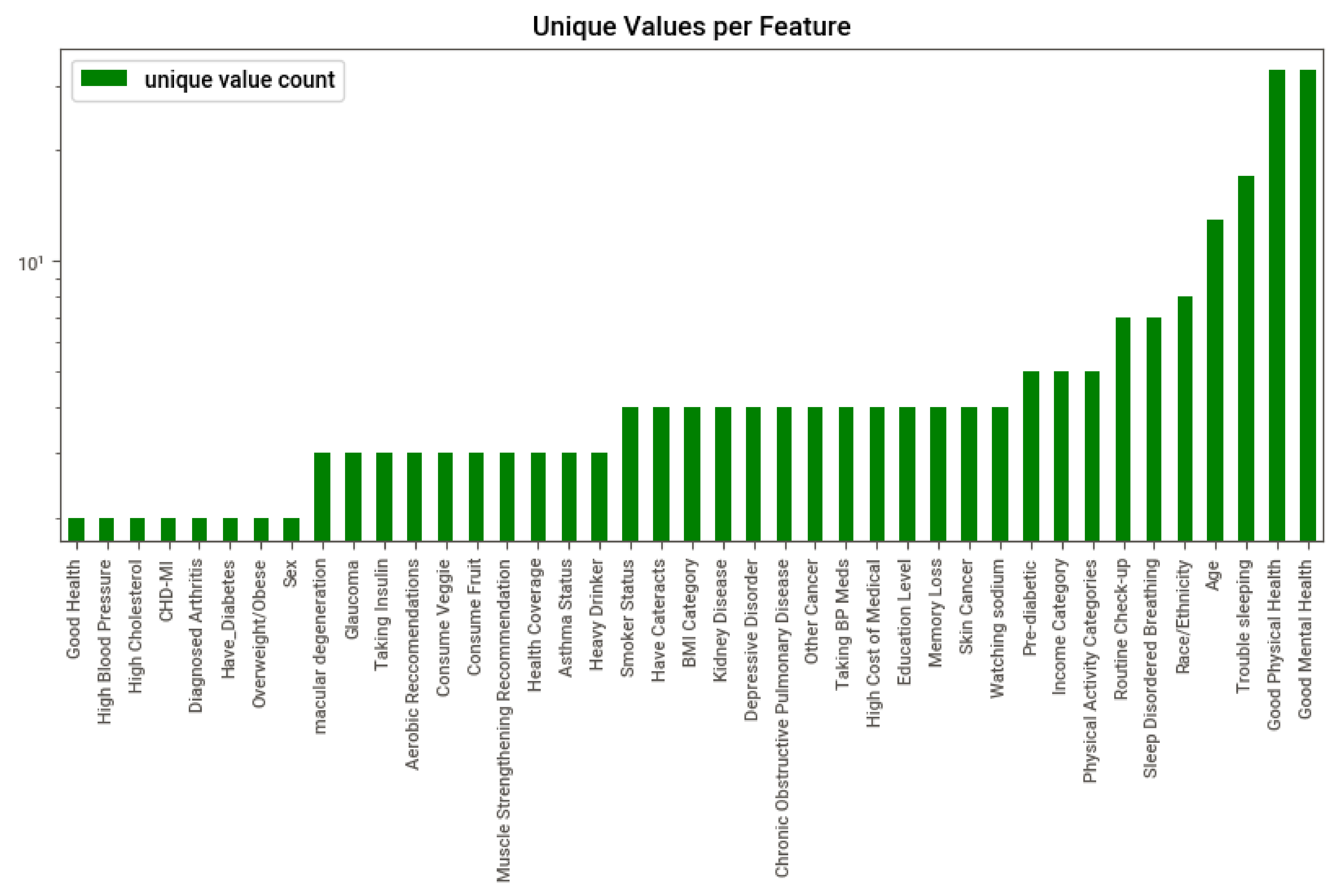
EDA - STRUCTURE INVESTIGATION

Shape

- 316,380 rows and 41 columns (including one output variable y)
- Converted columns to all numerical features

Unique Values

- The greatest unique values appeared in Good Physical Health and Good Mental Health
- A cluster trend



EDA - CONTENT INVESTIGATION

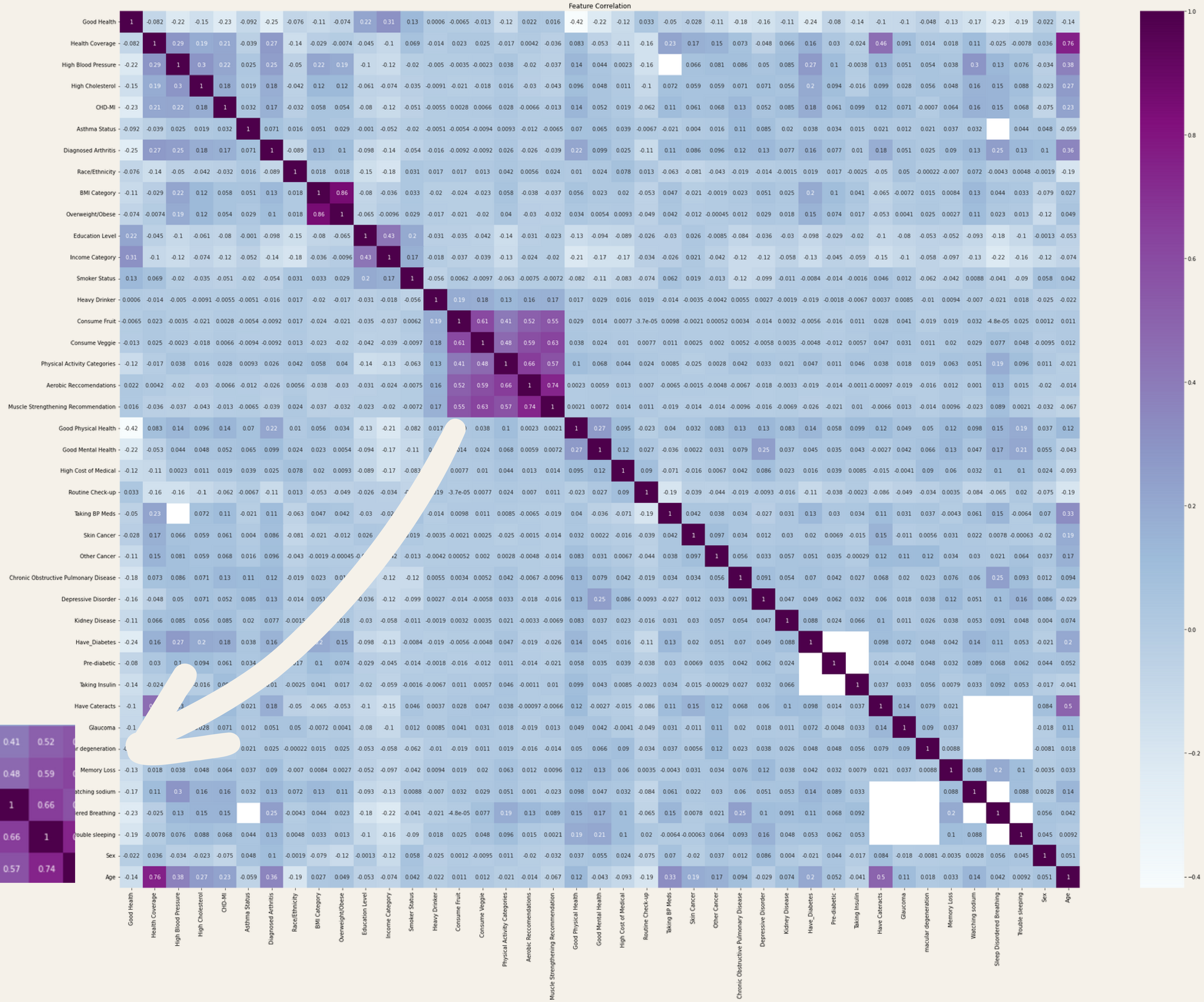
Feature Distribution

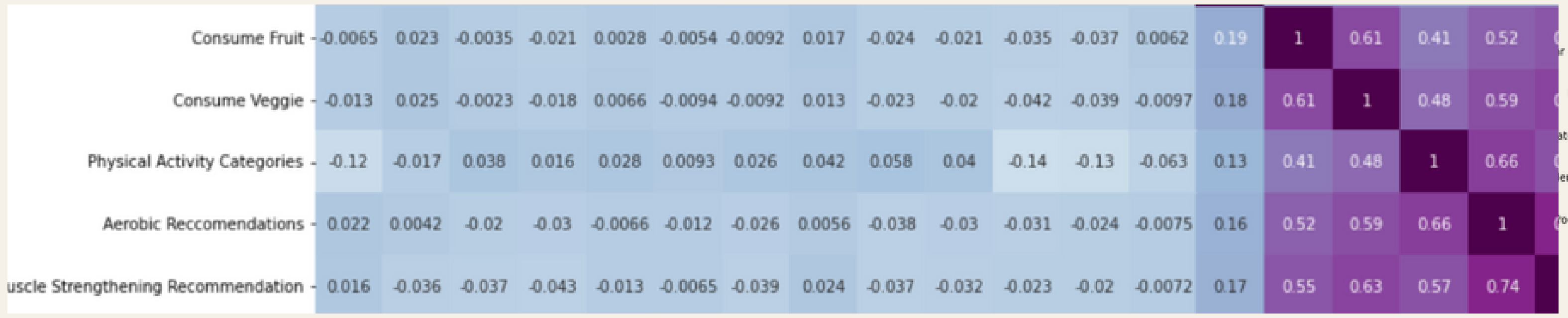
Most columns don't have much variability across values except features such as Age, Income Level, Education Level, BMI Category.

EDA - CONTENT INVESTIGATION

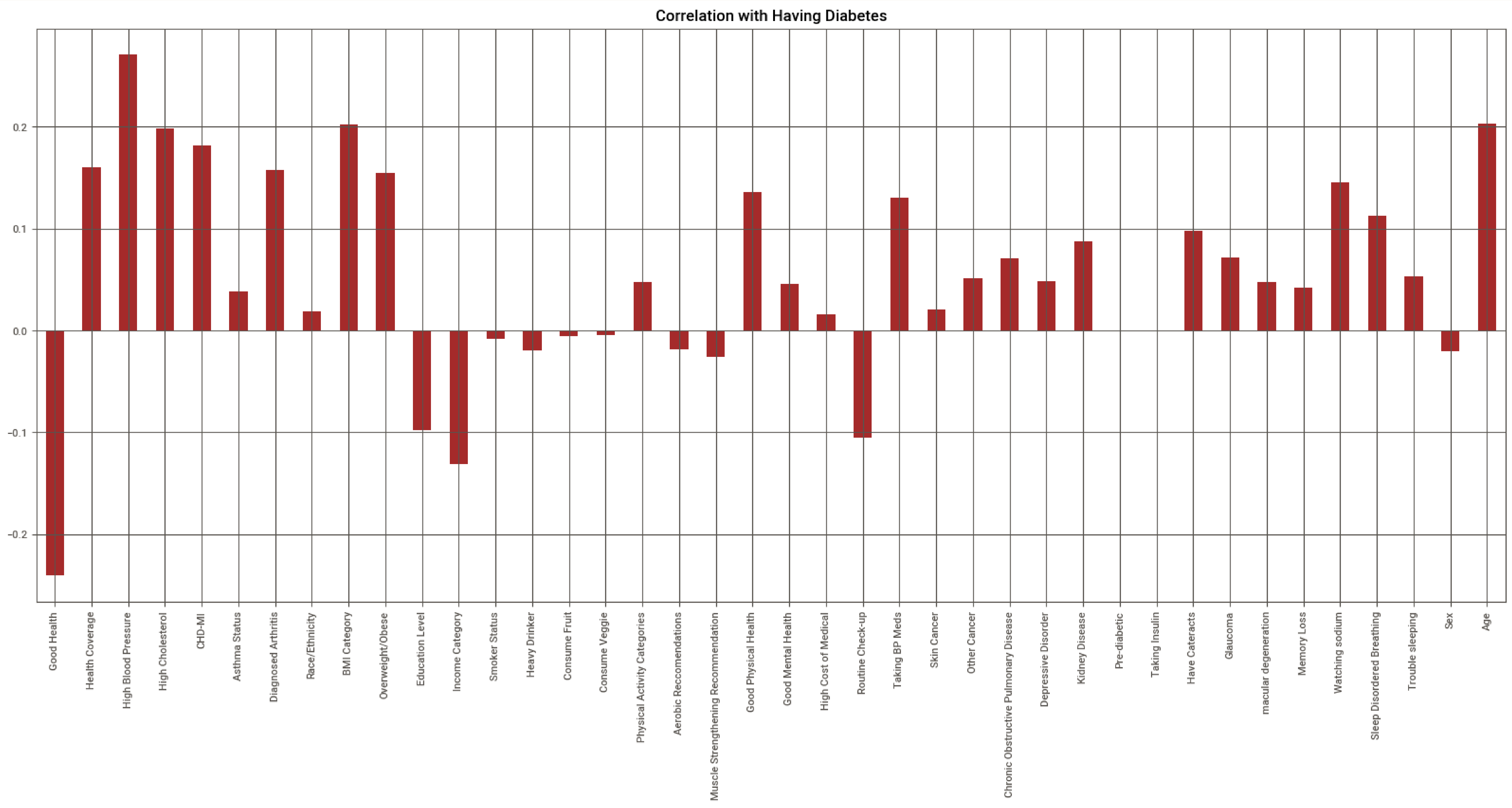
Feature Correlation

Strong correlations occur among features, including Consume Fruit, Consume Vegetables, Physical Activity Categories, Aerobic Recommendations, and Muscle Strengthening Recommendations.





EDA - CONTENT INVESTIGATION



Target Correlation

Most features are positively correlated to the target, albeit not significantly, except Good Health, Education Level, Income Category, Routine Check-up, and Gender.

REFERENCES

- Teboul, A. (2021, November 8). Diabetes health indicators dataset. Kaggle. Retrieved October 6, 2022, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv

- Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019, October 15). Predictive models for diabetes mellitus using machine learning techniques - BMC endocrine disorders. BioMed Central. Retrieved October 6, 2022, from https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6

- Diabetesjournals.org. (n.d.). Retrieved October 6, 2022, from https://diabetesjournals.org/care/article/36/12/3944/33144/Nonlaboratory-Based-Risk-Assessment-Algorithm-for

- Centers for Disease Control and Prevention. (2019, September 19). Building risk prediction models for type 2 diabetes using Machine Learning Techniques. Centers for Disease Control and Prevention. Retrieved October 6, 2022, from https://www.cdc.gov/pcd/issues/2019/19_0109.htm

- Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021, December 20). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review - diabetology & metabolic syndrome. BioMed Central. Retrieved October 6, 2022, from https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9

- Grandner, M. A., Seixas, A., Shetty, S., & Shenoy, S. (2016). Sleep Duration and Diabetes Risk: Population Trends and Potential Mechanisms. Current diabetes reports, 16(11), 106. https://doi.org/10.1007/s11892-016-0805-8

- https://www.cdc.gov/brfss/

Biology

THANK YOU!