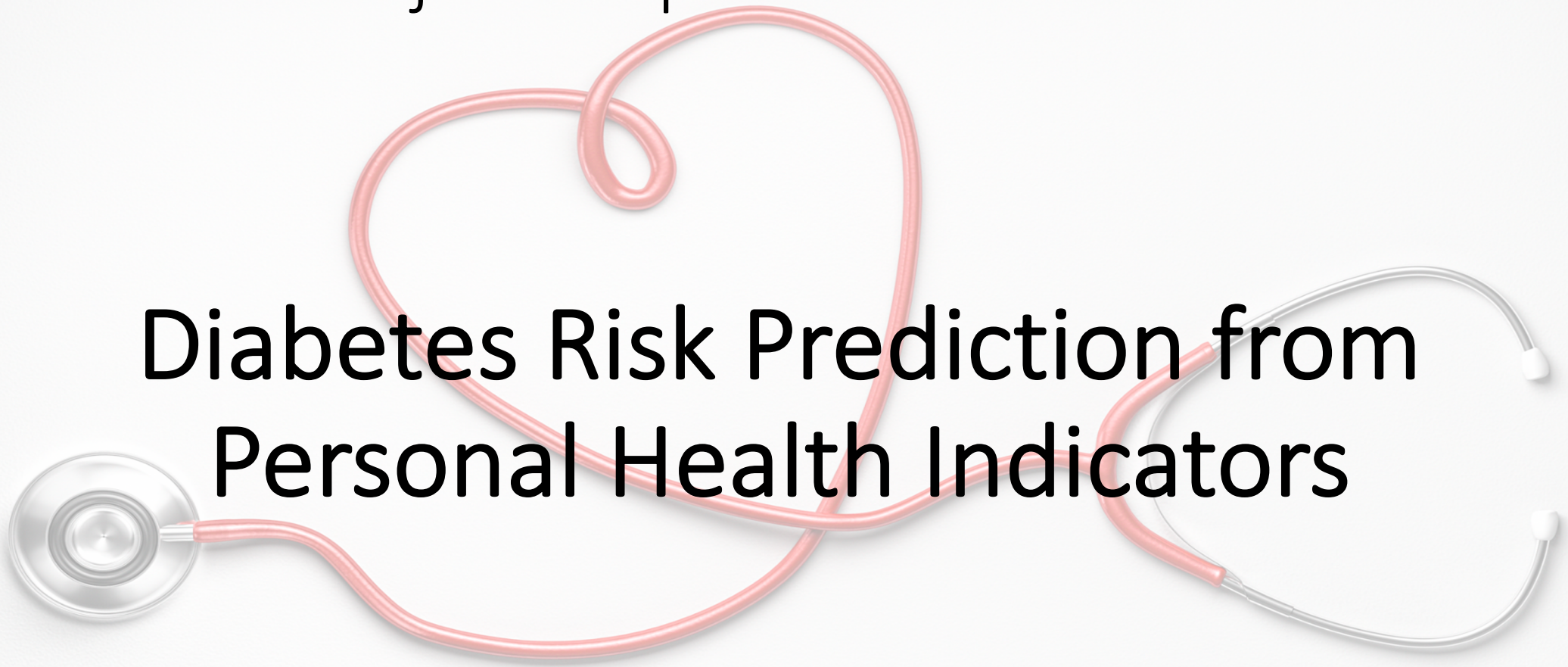
DSE-230
Project Proposal Presentation



Diabetes Risk Prediction from Personal Health Indicators

Team 3:
Chunxia Tong
Camm Perera
Sergey Gurvich

Diabetes: Background

Quick Facts:

- 37.3 million US adults have diabetes, and 1 in 5 of them don't know they have it.
- Diabetes is the seventh leading cause of death in the United States.
- Diabetes is the No. 1 cause of kidney failure, lower-limb amputations, and adult blindness.
- In the last 20 years, the number of adults diagnosed with diabetes has more than doubled.

Source: https://www.cdc.gov/diabetes/basics/diabetes.html

Risk Factors:

- Smoking: 19.8% were tobacco users based on self-report or levels of serum cotinine.
- Overweight and Obesity: 89.8% were overweight or had obesity, defined as a body mass
- Physical Inactivity: 34.3% were physically inactive
- Family history: Risk increases if a parent or sibling has type 2 diabetes
- Race or ethnicity: Black, Hispanic, American Indian and Asian American people — are at higher risk

Source: https://www.cdc.gov/diabetes/data/statistics-report/risks-complications.html

Health Complications:

- Heart disease
- Chronic kidney disease
- Nerve damage
- Other problems with feet, oral health, vision, hearing
- Mental health

Source: https://www.cdc.gov/diabetes/managing/problems.html

Problem Definition

Proposal: Diabetes Disease Risk Prediction Based on Personal Health Indicators

Source: https://www.lumahealth.io/

Possible Applications:

Medical

- Prevention
- Cost Reduction
- Treatment Improvement

Insurance

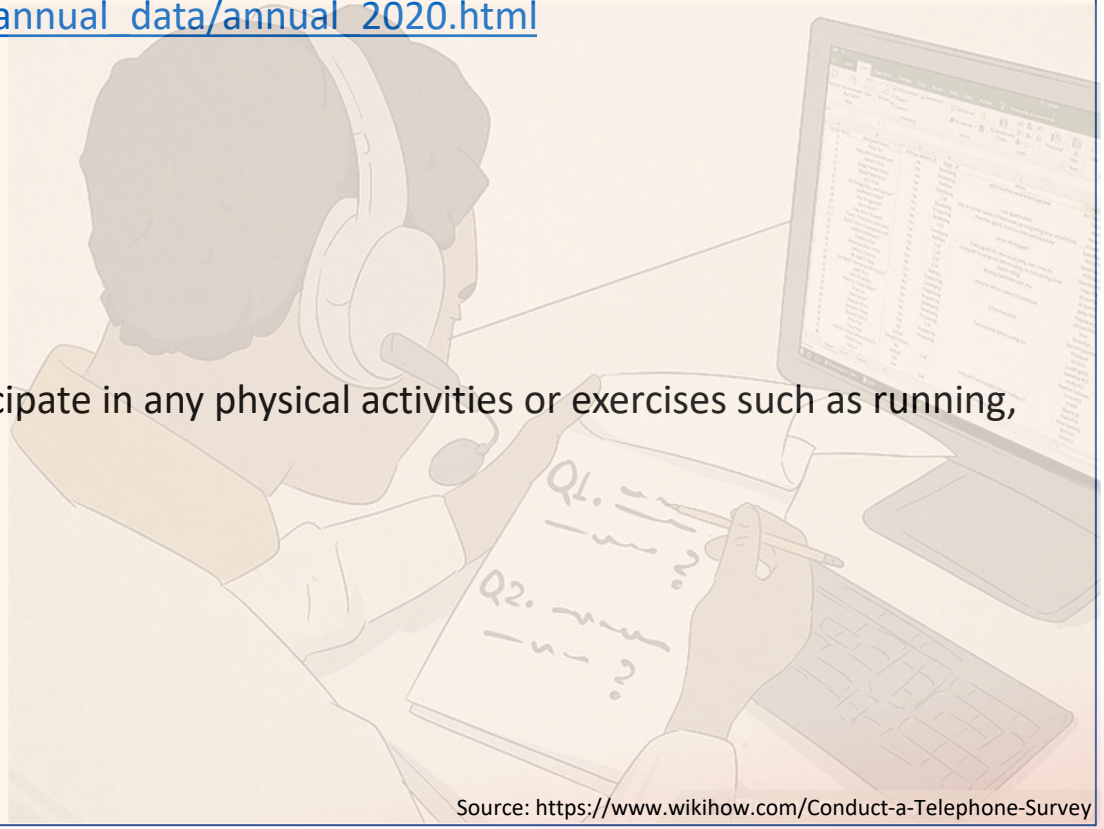
- Calculating Risks Adjustments for Medical Plans Pricing

Dataset Description

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

Source: https://www.cdc.gov/brfss/index.html

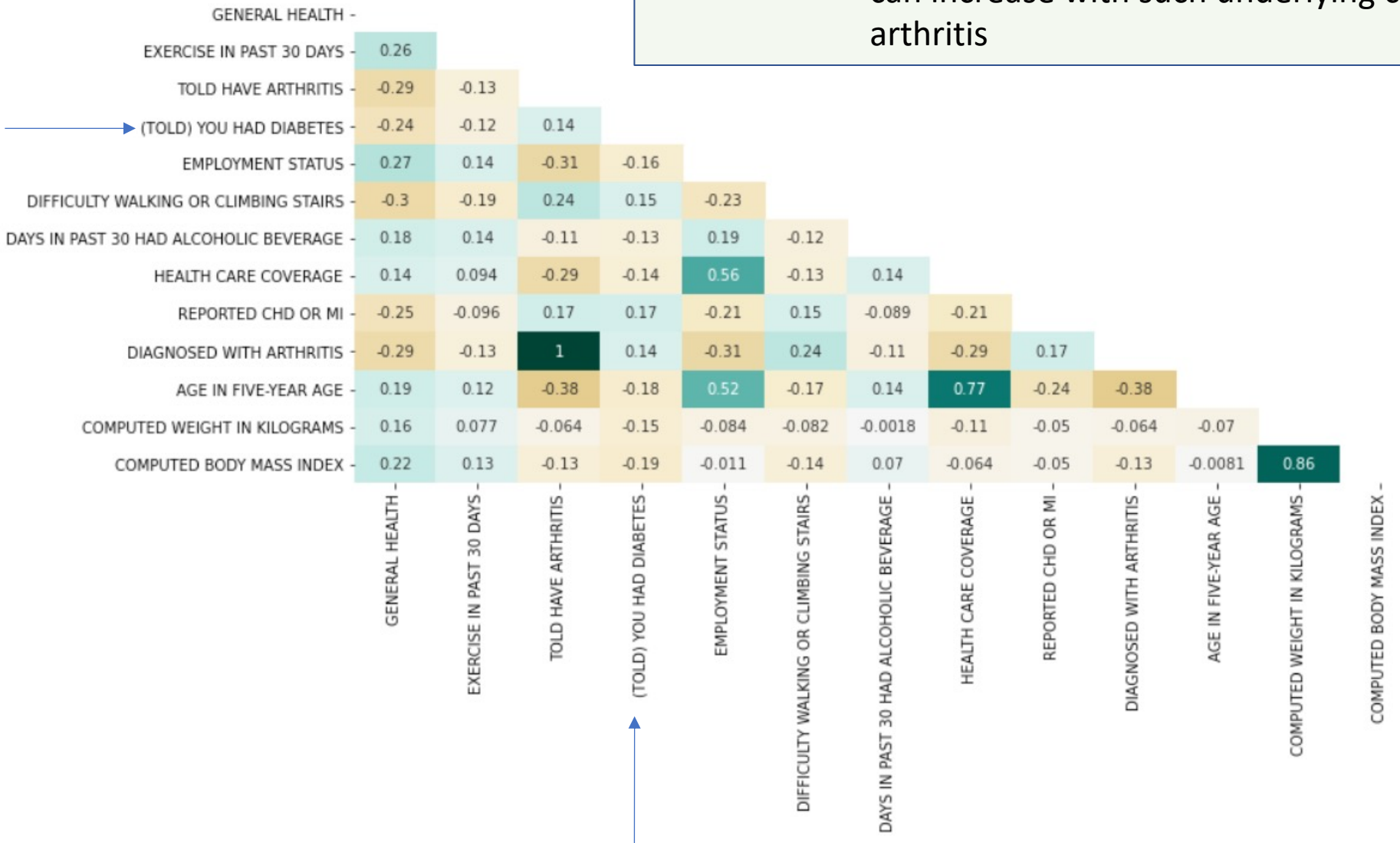
- Original Dataset Files, Scripts and Description: https://www.cdc.gov/brfss/annual_data/annual_2020.html
- Collection method – annual phone survey (2016-2020)
- Question Examples:
    - (Ever told) (you had) diabetes?
    - Do you now smoke cigarettes every day, some days, or not at all?
    - During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
- Calculated Variables Examples:
    - Calculated variable for fourteen-level age category (1-14)
    - Calculated variable for income categories (1-9)
    - Body Mass Index



Source: https://www.wikihow.com/Conduct-a-Telephone-Survey

Insights to be gained

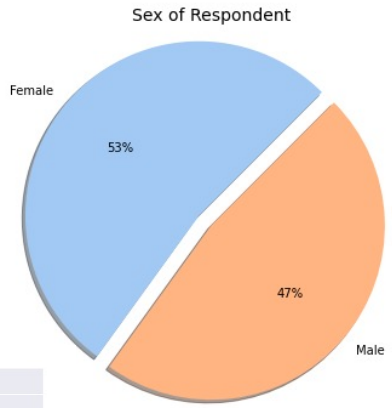
- Find a feature (or combination of features) that affect the chances of developing diabetes disease:
  - Example: diabetes risk level should decrease with exercise, but it can increase with such underlying conditions as heart disease or arthritis

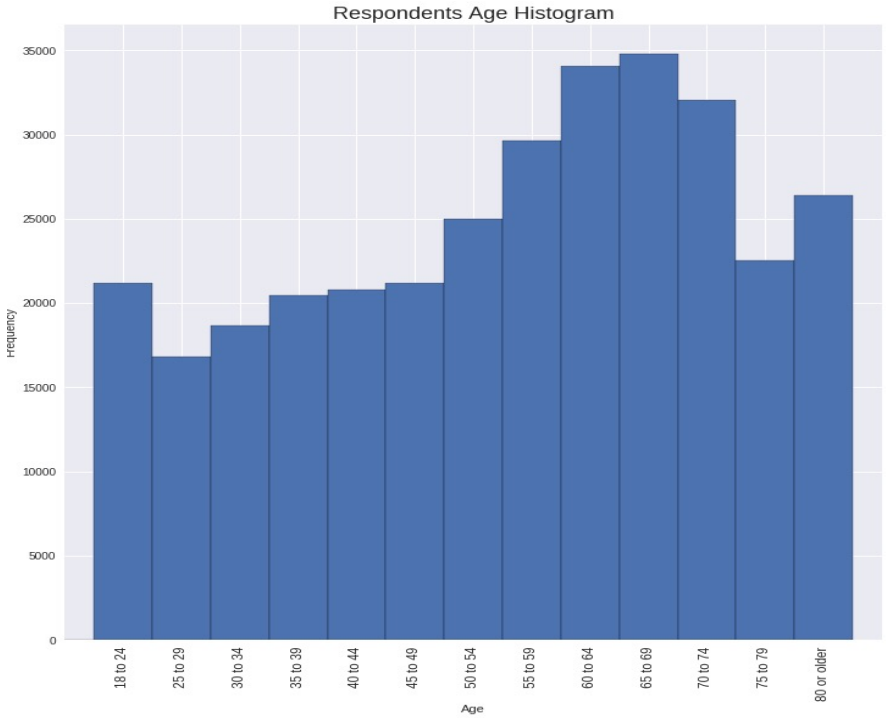


Analysis Task to Perform

EDA

- Data exploration
- Data visualization
- Summary statistics



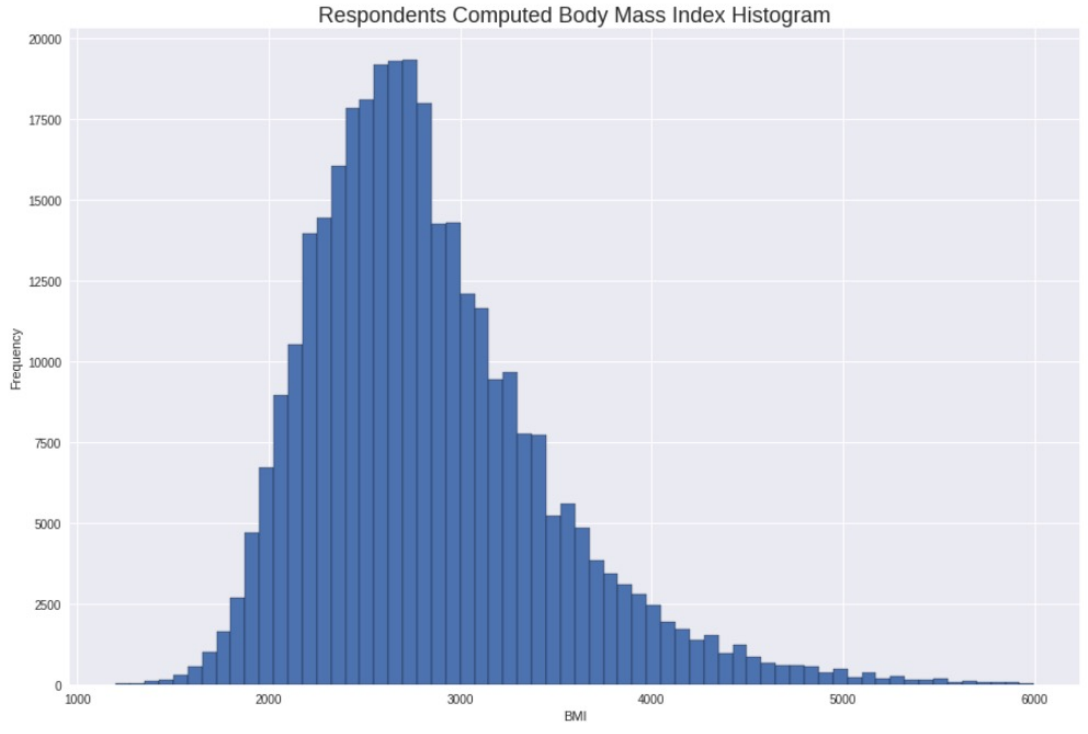


Data Preparation

- ✓ Non-sparse feature selection
- ✓ Remove rows with blank values (nulls/NAs)

(400k x 279 -> 327k x 112)

- Remove irrelevant features (Survey date, Phone number etc.)
- Remove non-contributing features/highly correlated features
- Remove rows with non-contributing answers ('refuse')
- Scaling
- Remove outliers
- Recoding Features



Modeling

- Classification (binary)

Potential challenges with data and/or task

- ✓ The provided data is in SAS format (data conversion)
- ❑ Correct features to select (feature reduction)
- ❑ Data engineering: Do we need to apply scaling/one-hot encoding?
- ✓ Which final target (disease) we predict
- ❑ Some of the correlations didn't make sense while EDA (at least commonsense)
- ❑ What model to choose?
- ❑ How do we define success?

Questions?

Thank you.