Modeling Mental Health & of Lifestyle Choices

Group 10:
Fatima Mora Garcia, fmora22@uic.edu, git: fmora22,
Zakareah Hafeez, zhafee3@uic.edu, git: zhafee3,
Shareek Shaffie, smoha45@uic.edu, git: shar33k,
Ikraam khan, ikhan55@uic.edu, git: ikraamkhan101,
Bader Rezek, breze2@uic.edu, git: baderrezek,
Subhi Kittaneh, skitt2@uic.edu, git: SKittaneh

Proposed Idea

The Idea

Investigating how lifestyle choices (e.g., physical activity, smoking, alcohol use) correlate with self-reported mental health outcomes

Why it matters

Mental health is a critical aspect of well-being, and understanding how lifestyle factors impact mental health can inform public health interventions and personal lifestyle adjustments

Research Question

How do lifestyle behaviors like exercise, smoking, and alcohol consumption impact mental health outcomes?

Hypothesis

We hypothesize that individuals who engage in regular physical activity, avoid smoking, and consume moderate amounts of alcohol report better mental health outcomes.

Data Cleaning Overview and Selected Features

DATA SUMMARY	
Data shape	433,323 rows x 350 columns
Columns	350
Rows	**433,323**
Missing values (by column)	**75,656,031**
CTELENM1	344,978
PVTRESD1	344,978
COLGHOUS	433,311

DATA SUMMARY	
Data shape	433,323 rows x 28 columns
Columns	28
Rows	**433,323**
Rows with missing values	383,061 (88.4%)
Duplicate rows	17,406 (4.0%)
Missing values (by column)	**1,234,358**

Dataset: BRFSS 2023 (Behavioral Risk Factor Surveillance System), surveying health behaviors and self-reported health outcomes across the U.S.

Dataset Size: ~433,323 rows, 350 columns.

Peek into Relevant Features:

- **Mental Health**:
  - MENTHLTH: Days of poor mental health in the past 30 days.
- **Lifestyle Choices**:
  - EXERANY2: Physical activity in the past 30 days.
  - SMOKE100: Whether the respondent smoked 100+ cigarettes in their lifetime.
  - SMOKDAY2: Current smoking status.
  - ALCDAY4: Days alcohol was consumed in the past 30 days.
  - AVEDRNK3: Average number of drinks on drinking days.
  - DRNK3GE5: Number of times the respondent had 5+ drinks in one sitting.
- **Demographics** (to control for confounding variables):
  - _AGEG5YR: Age in 5-year groups.
  - _AGE65YR: Two-level age category (18-64, 65+).
  - _RACEGR3: Race/ethnicity groups.
  - EDUCA, _EDUCAG: Education levels.
  - INCOME3, _INCOMG1: Income categories.

Challenges

Missing values (by column)	1,234,358
GENHLTH	4
PHYSHLTH	3
MENTHLTH	3
POORHLTH	181,153
EXERANY2	2
CHCSCNC1	3
CHCOCNC1	3
CHCCOPD3	3
DIABETE4	5
MARITAL	7
EDUCA	9
EMPLOY1	2,968
INCOME3	8,075
SMOKE100	19,674
SMOKDAY2	274,684
ALCDAY4	25,444
AVEDRNK3	221,197
DRNK3GE5	221,634
MAXDRNKS	222,037
FLUSHOT7	27,751
HIVTST7	29,613
_RACEGR3	86

Opening File

The dataset contained over 433,000 rows and 350 columns, which led to **performance issues** when loading and processing the file due to its large size.

Imputation

A significant number of columns were missing entries, and several features had **highly skewed** distributions which made median and mode imputation less effective

Balancing Data Quality

With large number of features, it was necessary to drop columns that had **excessive missing data and redundancy.**

Variable Encoding

A few categorical variables like _RACEGR# and INCOME#, required conversions using **encoding** techniques.

Mixed Data Types

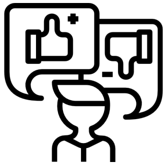
The dataset included a mix of numerical, categorical and binary data each requiring different **preprocessing techniques** increasing the **complexity to the workflow**

Data Exploration & Cleaning Continued

To ensure the dataset is ready for analysis, we applied the following data cleaning strategies to handle missing values but realized there where probably better approach because the data became too similar due to the large amount of missing values on some of the columns (advice from Professor on next slide):

- **Dropped Rows for Features with Minimal Missing Data**:
  - **Features**: GENHLTH, MENTHLTH, CHCSCNC1, CHCOCNC1, CHCCOPD3, DIABETE4, PHYSHLTH
  - **Reason**: These features had very few missing values , so we chose to drop the affected rows to avoid potential bias.
- **Imputed Missing Values for Features with Moderate Missing Data**:
  - **Categorical Features (Mode Imputation)**: SMOKE100, INCOME3, FLUSHOT7, HIVTST7
    - **Reason**: These features had a moderate percentage of missing values, and mode imputation was used to fill in the most frequent category.
  - **Numerical Features (Median Imputation)**: ALCDAY4, AVEDRNK3, DRNK3GE5
    - **Reason**: These features had moderate missing values, and median imputation was used as it's less sensitive to outliers than the mean.

Cleaning Cont.



After feedback from the Professor we have agreed on the following strategies to improve our data:

- **For the data that we want to use containing many missing values, we think it may be best to use random Coin Toss for Binary Missing Data:**
    - Assigned missing binary values (e.g., yes/no, smoker/non-smoker) using a random coin toss method. ***Purpose***: This ensures randomness when no clear trend is available for missing binary data.
- **Remaining Features of Relevance Not Yet Mentioned HERE**
- **Imputation** will be induced to allow for cohesion in the data set, any missing values being replaced with filler values allows for eros to be minimalized during computation
- We also plan on looking up more ways to fill in the null values in our data to reduce the amount of bias. We learned that any adjustment to the data will cause bias, but depending on the method we use, we may be able to get results that make more sense.

Next Steps

Feature Transformation:

- Normalize or scale numerical features.
- Apply one-hot encoding for categorical variables like EXERANY2 and MARITAL.

Initial Model Building:

- Build a simple linear regression model to predict mental health outcomes (MENTHLTH).
- Evaluate model performance using Mean Squared Error (MSE).

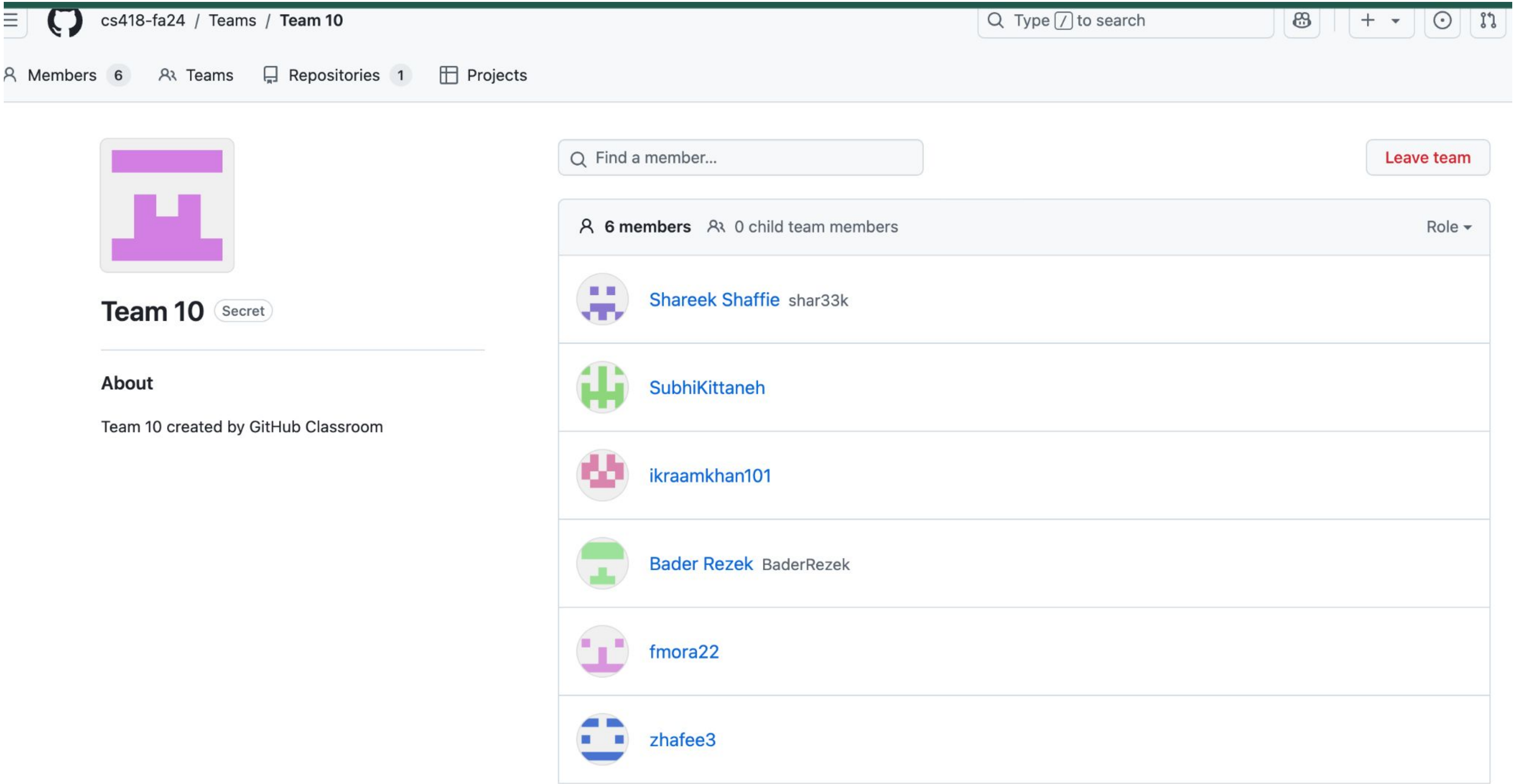
Handle Class Imbalance:

- Check for imbalance in MENTHLTH and address if needed (e.g., oversampling).

Visualize Insights:

- Visualize distributions of key variables (e.g., MENTHLTH, EXERANY2, SMOKE100).
- Analyze correlations between mental health and lifestyle factors.

Github


cs418-fa24 / Teams / Team 10
Type / to search
Members 6
Teams
Repositories 1
Projects
Find a member...
Leave team
Team 10
Secret
About
Team 10 created by GitHub Classroom
6 members
0 child team members
Role
Shareek Shaffie shar33k
SubhiKittaneh
ikraamkhan101
Bader Rezek BaderRezek
fmora22
zhafee3