Predicting Time to Diabetes Onset and Analyzing Diversity, Equity, and Equality in Research

Team #16

Yacine Marouf

Hunter Pozzebon

Priyonto Saha

...

CHL5230- APPLIED MACHINE LEARNING FOR HEALTH DATA

Fall 2023

University of Toronto

HIVE Lab Health Informatics, Visualization, and Equity

Institute of Health Policy, Management and Evaluation UNIVERSITY OF TORONTO

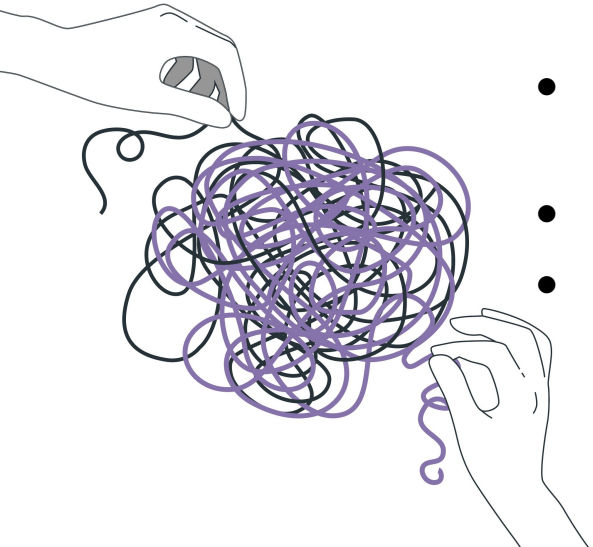
Dalla Lana School of Public Health

Research Questions and Dataset

- **Main Research Questions**
  - Can we predict the time to a diabetes onset?
  - What does current research about diabetes consist of?
    - How often is diversity and equity discussed in these papers?
    - What are the common methods used in research?
    - What are the common terms used in prevention methods?
- Diabetes prevalence is increasing at a rate of 3.3% per year. Predicting diabetes onset is important for prevention measures.
- Dataset is called Diabetes Study File 10K Dec 14 2017, from CPCSSM, with 10000 observations and 43 features.

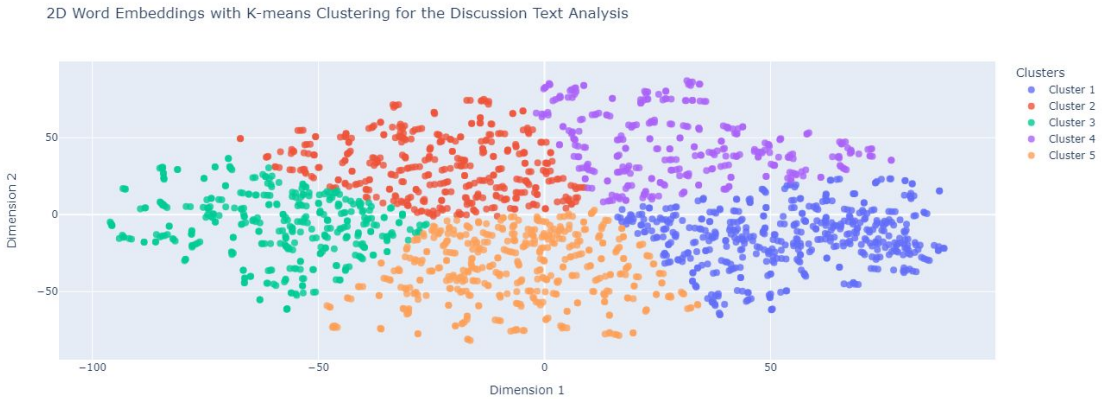
Methodology (Inside the Box)

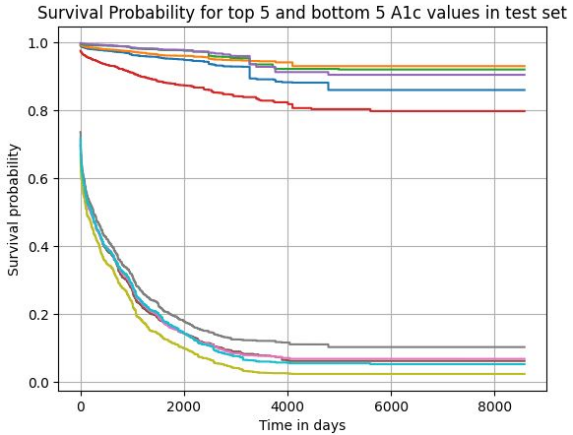
- **Random Survival Forest (RSF)** to predict time to diabetes onset
  - Data is primarily composed date/time features
  - Accommodates right censoring
- **GloVe** and **K-Means Clustering** for analysis of research
  - GloVe chosen for global text context
  - K-Means chosen for unlabeled text data exploration
- **Core features**: Survival Time, Comorbidities, Biomarkers, Manuscript Text.
- **Target variable**: Time to Diabetes onset for RSF
- **Survival analysis methods** and **impacts of bias and diversity** unexplored in current literature.



Results

- 100 trees trained with max depth of 15, min sample split of 150, and min sample leaf of 100
- Random Survival Forest Model resulted in a **high concordance index of 0.83**
- **A1c** and **fasting blood sugar (FBS)** are the most important for prediction
- K-means showed diversity terms were closely related to systems, structures, and institutions





Novel Method for Predicting Time to Diabetes Onset

- RSF model has high performance and potential for diabetes screening and prevention
- **Limitations**
  - Left censored data
  - Computational intensity
- **Findings within text analysis**
  - Stigmatized language did not appear in the word cloud
  - More needs to be done in relating populations to the data analysis