Predicting Lifestyle Disease in the Canadian Population

Carleton University

Genevieve Forget
Psychology
genevieveforget@cmail.carleton.ca

Nicholas Pontone
Geography
nicholaspontone@cmail.carleton.ca

Chandra Kotillil
MBA
chandrasekharkotill@cmail.carleton.ca

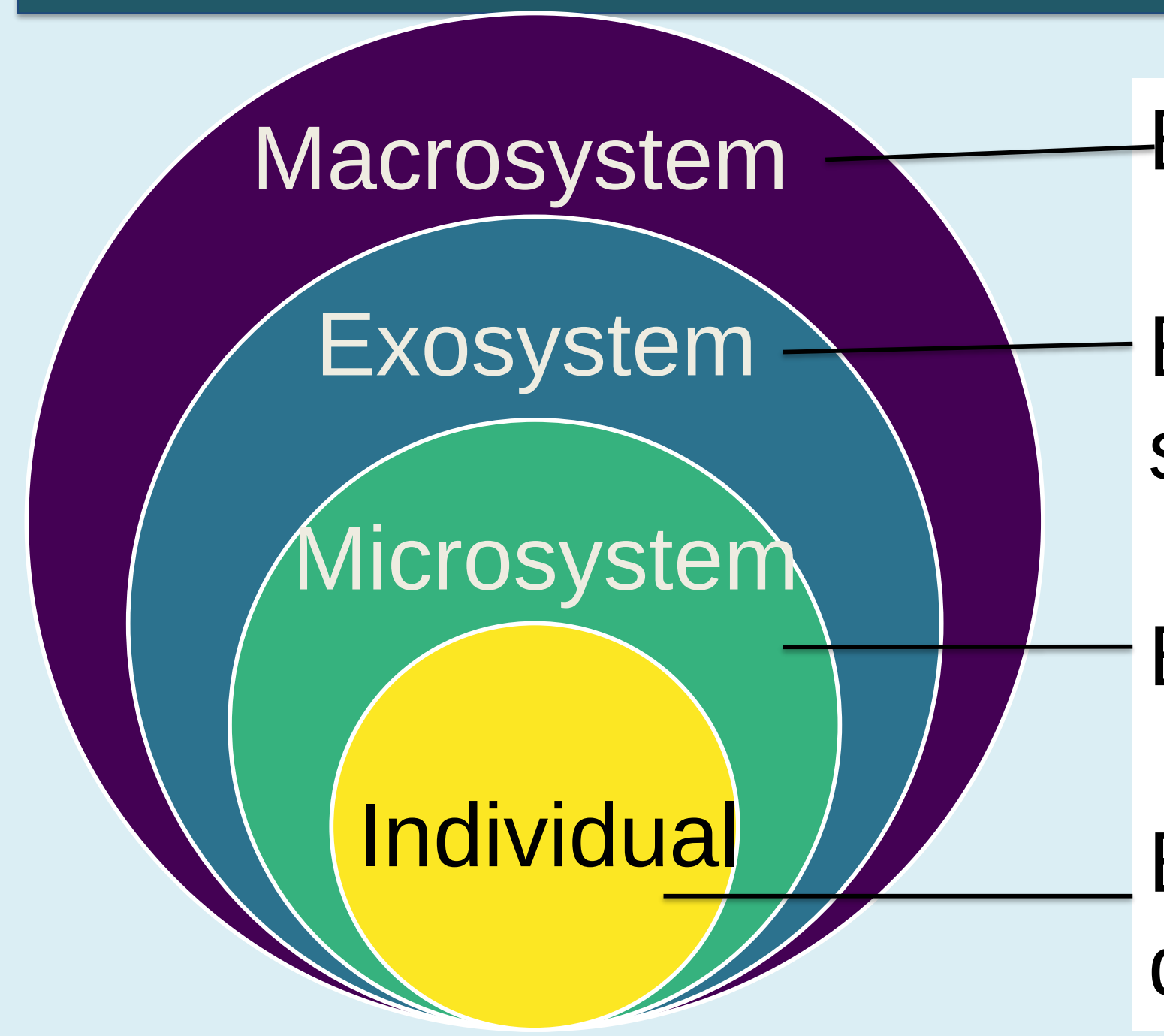
Tuheen Ahmmed
Computer and Electrical Engineering
tuheenahmmed@cmail.carleton.ca

Majid Komeili
Supervisor
MajidKomeili@cunet.carleton.ca

Background

- Cardiovascular disease (CVD) and Type 2 diabetes (T2D):
  - Are among the top 10 causes of death in Canada [1]
  - Pose a significant burden on the Canadian economy: annual cost of CVD is approx. $21.2 Billion, while for T2D is just under $30 Billion [2,3]
  - CVD prevalence in Canada remains stable at 8.9% [4]; hospitalizations for structural heart disease increased by 50% from 2007 to 2017 [2]
  - T2D prevalence was estimated at 9.3% of the Canadian population in 2015 and is predicted to rise to 12.1% by 2025 [5]
- Individual, social, and environmental factors have been linked with an increased risk of CVD and T2D [6]

Bioecological Theory [7] – A Simplified Model



E.g., culture, environment

E.g., access to health services

E.g., friends, family

E.g., health, behaviour, demography

Objectives

- To study to what extent machine learning classifiers can identify individuals who are 0 = *Healthy*, individuals diagnosed with 1 = *CVD*, 2 = *T2D* , 3 = *Both*
- To identify important features of prediction to formulate evidence-based recommendations for the prevention of CVD and T2D

Data

Canadian Community Health Survey – Annual (2018) [8,9]

- Health, social, demography, and economy

Environment and Climate Change Canada, Canadian Forest Service

- Air quality and climate normals, % canopy cover [10,11,12]

Methodology

Data Preprocessing:

- Environmental data interpolated using Empirical Bayesian Kriging
- Null values were removed

Classification:

- 70:30 stratified training and validation split
- 29 important features were subset from a selection of 113 features with theoretical basis
- Adaptive Boost Classifier (AdaBoost) with 3000 iterations
- Random Forest Classifier with 3000 trees

Next Steps:

- Compare most important predictors derived from theory vs those decided using feature selection algorithm
- Testing and comparing other variable subsets (i.e., health behaviour, social stress theory[13])

Results

Highly imbalanced class data

-	Healthy	=	97417 (86.39%)
-	CVD	=	8175 (7.25%)
-	T2D	=	5243 (4.65%)
-	Both	=	1928 (1.71%)

Figure 1. Correlation Heatmap

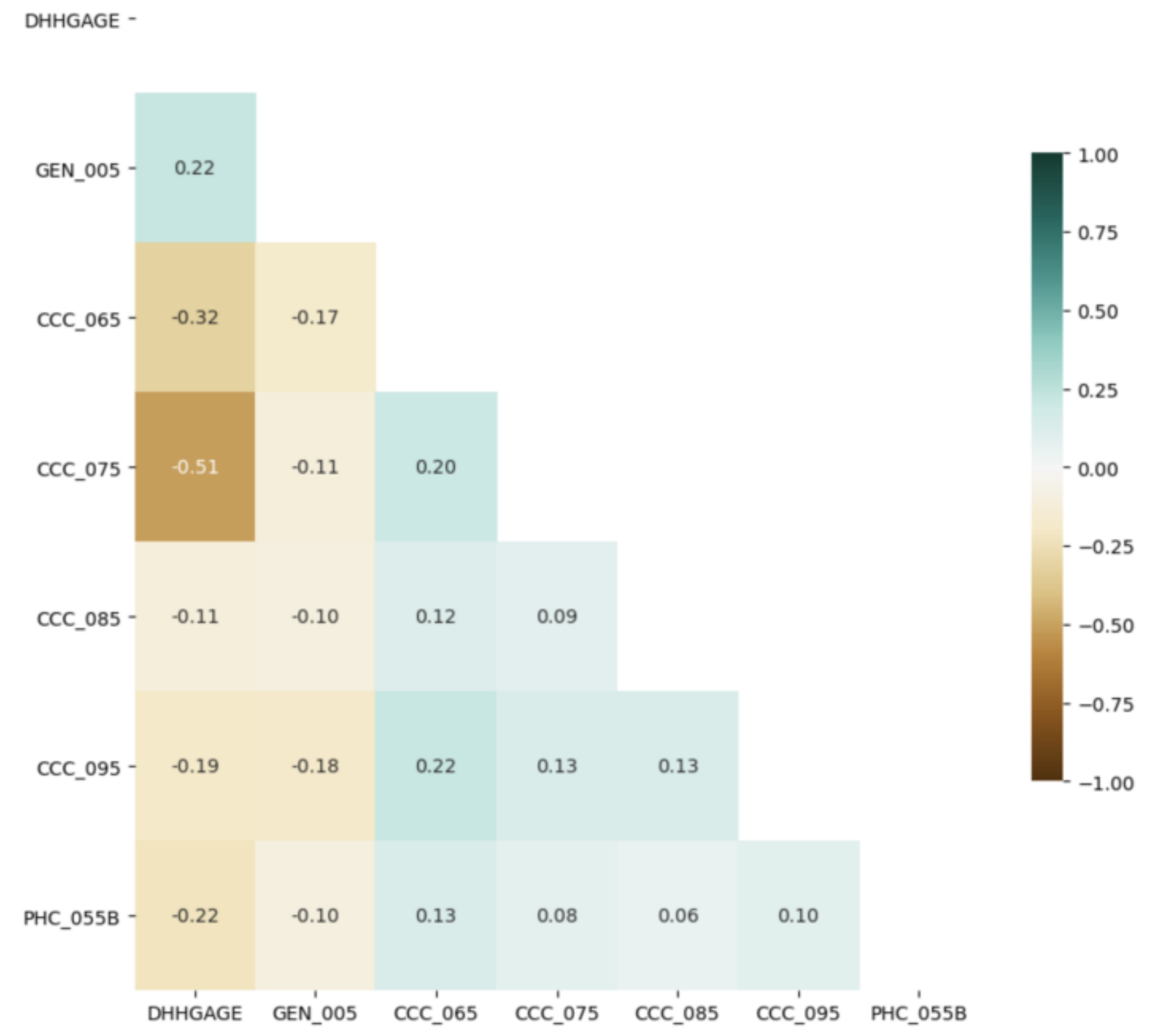


Figure 2. AdaBoost Confusion Matrix

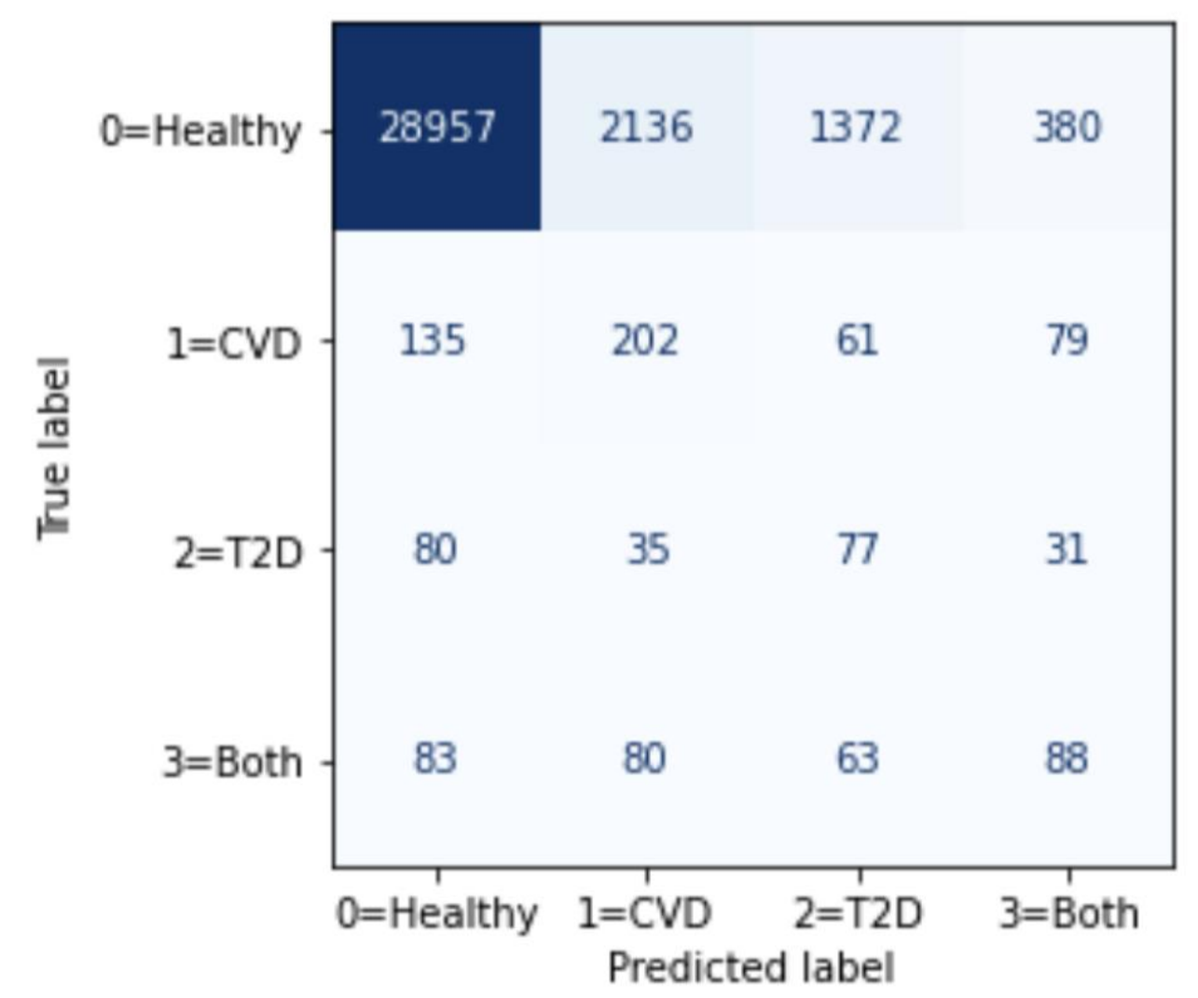
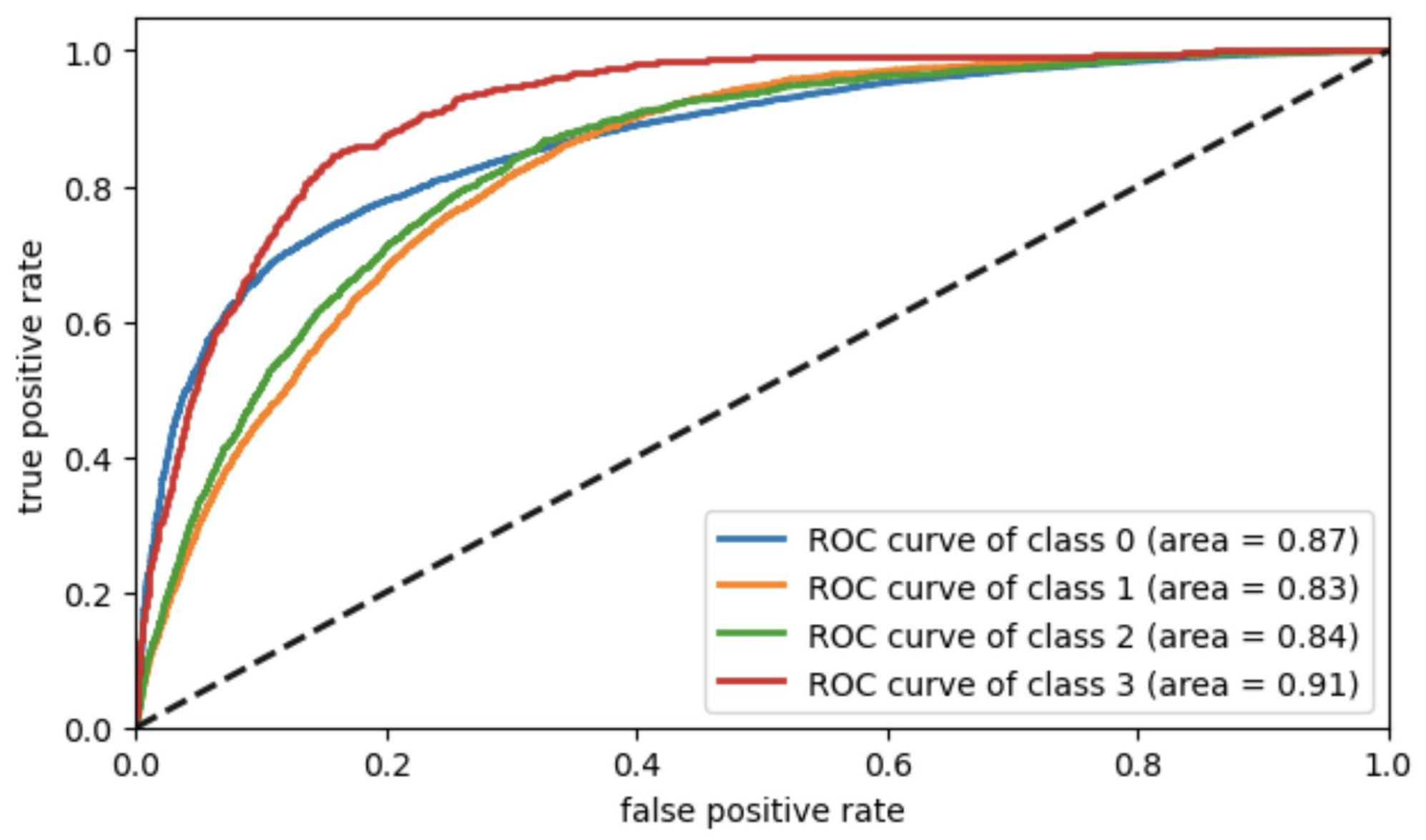


Figure 3. AdaBoost ROC Curve



AdaBoost w/ 3000 iterations

- Overall accuracy of 86.6%
- Log loss of 1.38

Random Forrest w/ 3000 trees

- Overall accuracy of 86.5%
- Log loss of 0.415

Feature selection comparison

- Automated feature selection using percentile (F1 score ANOVA) performs better than handpicked features

Figure 4. AdaBoost Precision-Recall Curve

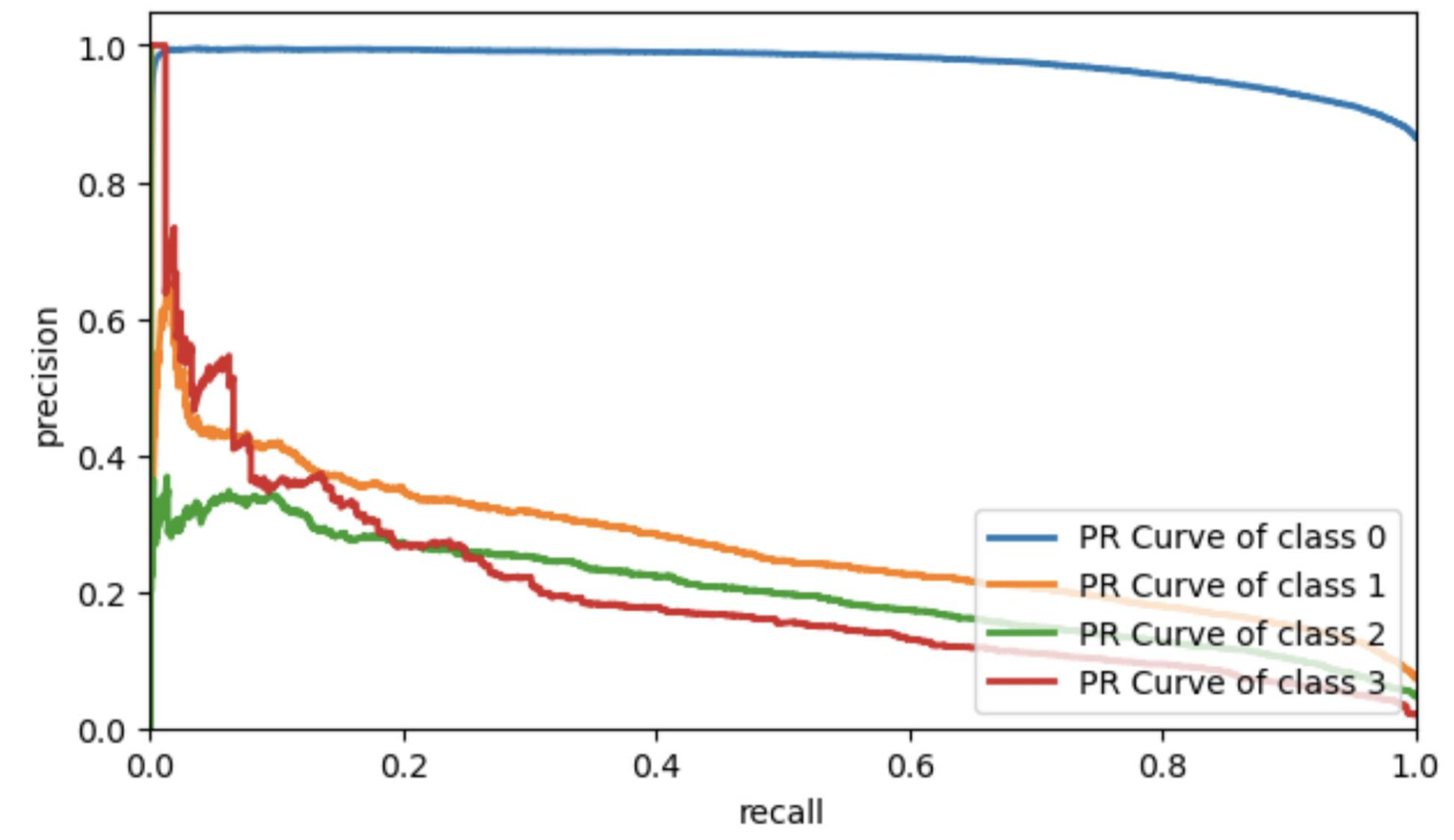


Figure 5. Biecological Theory Features

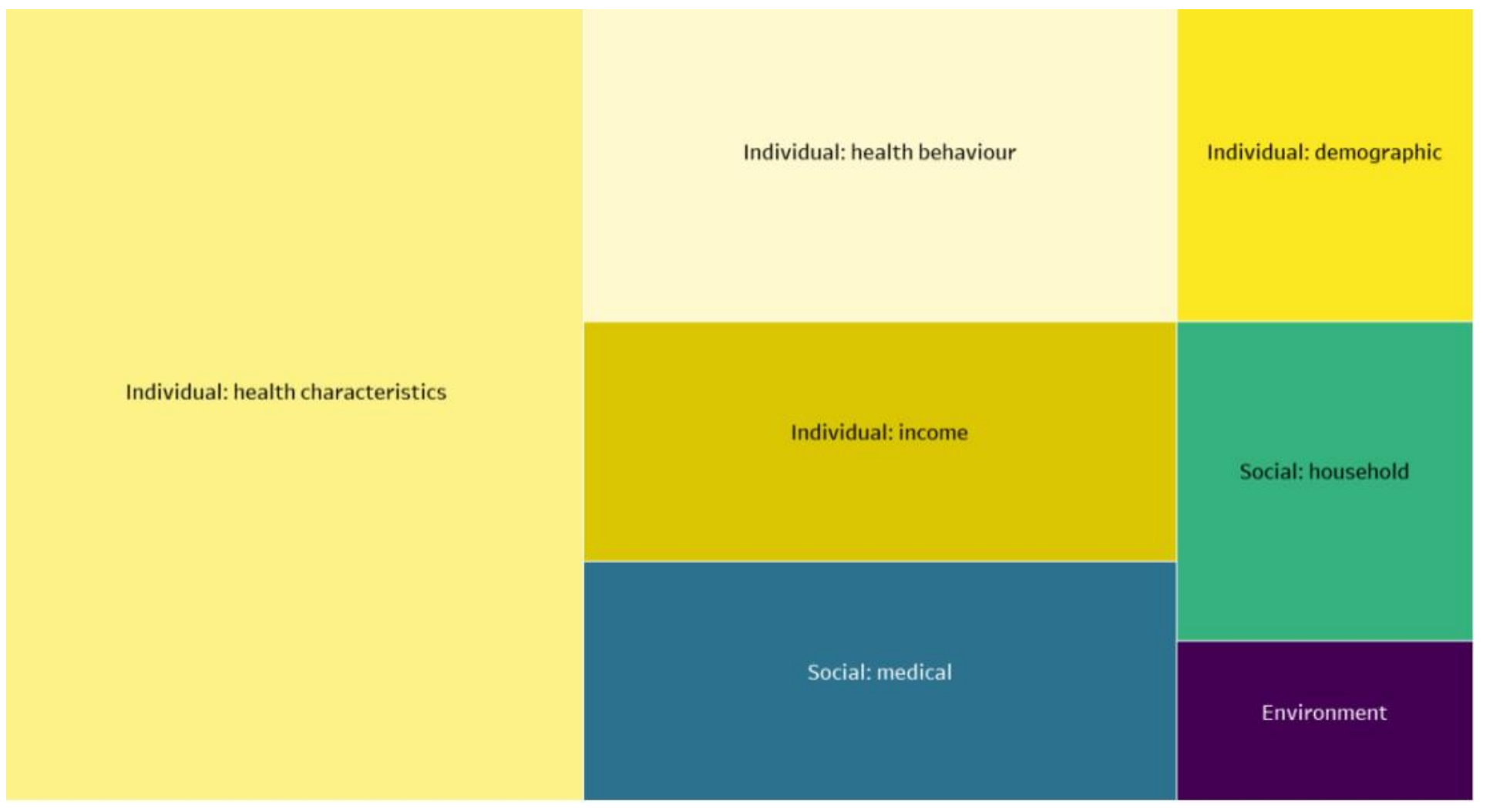
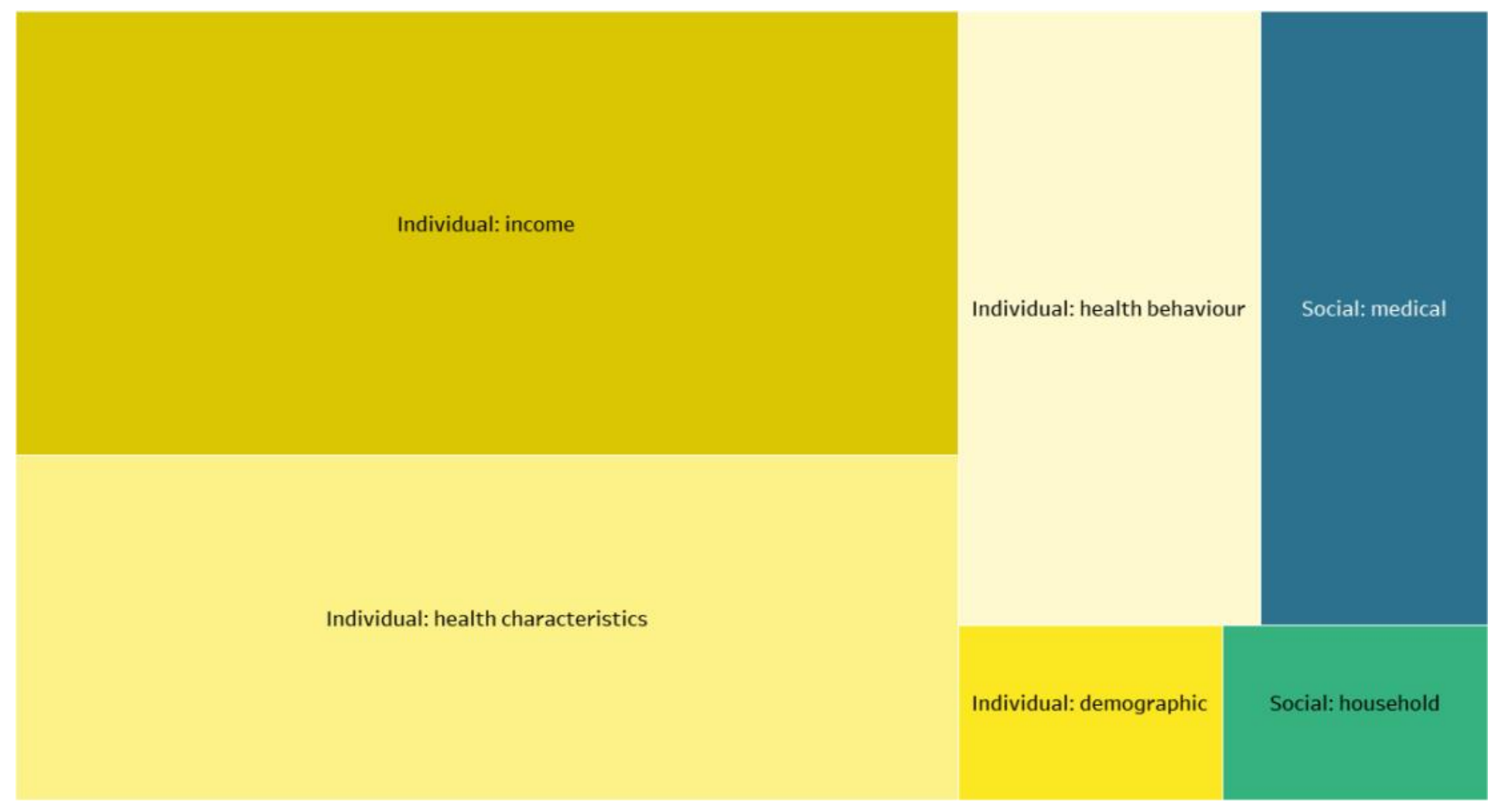


Figure 6. Percentile Selected Features



Conclusion

Descriptive

- No strong correlation between top features
- Positive correlation between blood pressure and age

Predictive

- Able to predict occurrences of CVD, T2D, and both.
- High rate of false positives, few false negatives

Prescriptive

- Can be used as a pre-screening tool, or to identify those at risk of developing CVD and/or T2D
- Modifiable risk factors: sedentary behaviour, smoking, and alcohol consumption patterns

Limitations and Future Directions

- Environmental data was not an effective predictor at the health region scale. Future work should include finer grain geographic data (i.e., postal code level)
- Data are cross-sectional; future studies should use longitudinal data to establish temporal precedence

References

[1] Statistics Canada, “Leading causes of death, total population, by age group.” Government of Canada. doi: 10.25318/1310039401-ENG. **[2]** Canada by the Numbers. (2019).Heart & Stroke Foundation of Canada. Retrieved 24 March 2022, from https://www.heartandstroke.ca/articles/connected-by-the-numbers **[3]** New Data Shows Diabetes Rates And Economic Burden On Families Continue To Rise In Ontario. (2022). Diabetes Canada. Retrieved March 24, 2022, from https://www.diabetes.ca/media-room/press-releases/new-data-shows-diabetes-rates-and-economic-burden-on-families-continue-to-rise-in-ontario **[4]** Heart Disease Canada. (2018). Government of Canada. Retrieved March 24, 2022, from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html. **[5]** Chapter 1: Introduction. (2022). Diabetes Canada. Retrieved 25 March 2022, from https://www.diabetes.ca/health-care-providers/clinical-practice-guidelines/chapter-1 **[6]** Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, “Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review,” Am. J. Prev. Med., vol. 61, no. 4, pp. 596–605, Oct. 2021, doi: 10.1016/j.amepre.2021.04.016. **[7]** U. Bronfenbrenner, “The Ecology of Human Development: Experiments by Nature and Design,” Harvard University Press, 1979. **[8]** Y. Beland, “Canadian Community Health Survey — Methodological overview,” Health Rep., vol. 13, no. 3, p. 6, 2002. **[9]** ”Canadian Community Health Survey - Annual Component (CCHS) 2017-2018”, Statistics Canada, Jan. 2020. [Online]. Available: https://hdl.handle.net/11272.1/AB2/SEB16A **[10]** “National Air Pollution Surveillance (NAPS) Program”, Environment and Climate Change Canada, Aug. 2019. [Online]. Available: https://www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/national-air-pollution-program.html **[11]** "Canadian Climate Normals”, Environment and Climate Change Canada, Jan. 2020. [Online]. Available: https://climate.weather.gc.ca/climate_normals/ **[12]** Beaudoin, A., Bernier, P.Y., Villemaire, P., Guindon, L., Guo, X.-J. 2017. Tracking forest attributes across Canada between 2001 and 2011 using a kNN mapping approach applied to MODIS imagery, Canadian Journal of Forest Research 48: 85–93. DOI: https://doi.org/10.1139/cjfr-2017-0184 **[13]** Pearlin LI. The sociological study of stress. Journal of Health and Social Behavior. 1989;30:241–256. https://doi.org/10.2307/2136956