**Revised Project Abstract: An Exploratory Analysis of Diabetes and Related Health Indicators**

Stephanie Boiossonneault

CIND820 D1H

Supervisor: Tamer Abdou, PhD

October 06, 2024

Good health and well-being is one of the 17 Sustainable Development Goals in the Government of Canada’s 2030 Agenda (Government of Canada, 2024b). To help achieve this goal, the Government of Canada works towards preventing non-communicable diseases, the leading cause of premature deaths globally (Government of Canada, 2024a). Diabetes is a widespread non-communicable disease affecting 1 in 10 adults worldwide (World Health Organization, n.d. as cited in Statistics Canada, 2023). There are different types of diabetes, each posing several potential health complications including “kidney disease, foot and leg problems, eye disease (retinopathy) that can lead to blindness, heart attack & stroke, anxiety, nerve damage, amputation and erectile dysfunction” (Diabetes Canada, n.d.). This research project proposes a secondary exploratory analysis of publicly available Centers for Disease Control (CDC) health survey data to gain a deeper understanding of diabetes health indicators.

The proposed dataset for this project is a consolidated version of the 2014 Behavioral Risk Factor Surveillance System (BRFSS) Survey titled “diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015.csv” (Teboul, 2021; UC Irvine Machine Learning Repository, 2023). It consists of 22 variables and a sample of 253,680 observations from the CDC’s annual health-related telephone survey examining risk behaviors, chronic health conditions, and use of preventative services from 400,000 respondents across 50 states in the US (Teboul, 2021; Centers for Disease Control, 2015). The dataset contains 6 numeric type and 16 categorical type variables, one of which is Diabetes\_012 classifying respondents as either having no diabetes or only diabetes during pregnancy, prediabetes, or diabetes. This project proposes using tools such as Jupyter Notebook, Visual Studio Code, Python programming language, and Python libraries to apply data mining and knowledge discovery techniques to the data to uncover patterns and trends between diabetes diagnosis and other health-related factors.

Research questions related to the project have the goal of helping to identify individuals living with diabetes and understand their needs, as this is becoming increasingly important for supporting the refinement of current screening tools and programs designed to help support Canada’s goal of health and well-being. One research question to consider is “What sets of health conditions and lifestyle choices commonly occur together in individuals diagnosed with prediabetes and diabetes?”. Applying knowledge induction using the apriori algorithm to generate sets of association rules will help answer this question and help with designing programs that inform and target change towards diabetes and associated heath conditions and lifestyle choice’s. Other questions to consider are “What machine learning model has the best performance measures towards successfully identifying a diabetes diagnosis?” and “What features are the highest predictors of a diabetes diagnosis based on that model?”. This project therefore proposes applying several supervised learning techniques including classification decision tree, random forest, multinomial logistic regression, K-Nearest Neighbors, and Naive Bayes, followed by an evaluation of their generalization capability to explore and select the best machine learning models capable of ability to predicting or identifying diabetes diagnosis based on demographics and health-related factors. Identifying features important to the diagnosis of diabetes can inform improvements in screening techniques used by clinicians to identify the disease. Lastly, answering the question “Are there any clusters of demographic groups, dietary behaviors, or healthcare access issues that exist among individuals with different diabetes diagnoses?” can provide further insight into Canadian’s experience of diabetes. Applying K modes and K means clustering will help answer this question and guide the development of successful policies and programs that further address Canadian needs through policies and programs to help reduce diabetes prognosis.

**References**

Centers for Disease Control and Prevention. (2015, September). Behavioral risk factor surveillance system: Overview BRFSS 2014. <https://www.cdc.gov/brfss/annual_data/2014/pdf/Overview_2014.pdf>

Diabetes Canada. (n.d.). What is diabetes?. <https://www.diabetes.ca/about-diabetes-(3)/what-is-diabetes>

Government of Canada. (2024a, June 1). Sustainable development goal 3: Good health and well-being. https://www.canada.ca/en/employment-social-development/programs/agenda-2030/health-well-being.html

Government of Canada. (2024b, June 17). Canada and the sustainable development goals. <https://www.canada.ca/en/employment-social-development/programs/agenda-2030.html#sdg>

Statistics Canada. (2023, November 29). Diabetes among Canadian adults. <https://www.statcan.gc.ca/o1/en/plus/5103-diabetes-among-canadian-adults>

Teboul, A. (2021). Diabetes health indicators dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

UC Irvine Machine Learning Repository. (2023, September 9). CDC Diabetes Health Indicators. <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>