JANUARY 25, 2025

CIND820: CAPSTONE PROJECT

PROJECT ABSTRACT

ASMA SHAIKH
052129962

A. CONTEXT

In 2021, International Diabetes Federation (IDF) reported approximately 10.5% of the adult population from age 20 to 79 has diabetes. Almost half of patients are unaware of their diagnosed condition and are living without any awareness and cautionary measure. It is estimated by IDF that by 2045 there will be 1 in 8 adults living with diabetes. Diabetes is expected to more than double by 2050[1]. Diabetes can cause long term damage to human function including the following but not limited to blindness, heart attacks kidney failure, and stroke according to World Health Organization article on Diabetes dated 14 November 2024[2]. Diabetes data from the National Health and Nutrition Examination Survey revealed that during 2021- 2023, the total diabetes case was 15.8%, of which 4.5% were undiagnosed diabetes adults from United Status[3].

B. PROBLEM STATEMENT

The intent is creating a machine learning model to predict diabetes based on key categories based on their health, demographics and behavior parameters. Some of the research questions are:

1. what are major factors and correlations that increase the likelihood of diabetes?
2. Does gender or age increase the probability of diabetes?
3. What is the relationship between individual Body Mass Index (BMI) and positive diabetes?
4. How smoking history related to diabetic and non-diabetic individuals
5. Is there any relationship between Hemoglobin A1c level and blood glucose level?
6. Which machine learning model accurately predicts diabetes?

C. DATASET EXPLORATION

The dataset (#1) is a collection of health indicators and demographics and behaviors data from patients, along with binary classification diabetes status (No Diabetes, Diabetes). There are 100,000 rows in this dataset. The health indicators includes bmi, HbA1c level, blood glucose level

and demographics and behaviors included gender, age, smoking history. The source of the dataset is Electronic Health Records (EHRs) and download from this link: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

The dataset (#2) is a collection of health indicators and demographics data from patients, along with three categories Diabetes status (0 = No, 1 = Prediabetes, 2 = Diabetes). There are 253,680 rows in this dataset. The health indicators includes highbp, highchol, bmi, smoker, stroke, etc. and demographics and behaviors included physical activity, fruits, veggies, sex, age, etc. The source of the dataset is Behavioral Risk Factor Surveillance System (BRFSS) and download from this link: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

D. TOOLS &TECHNIQUES

The purpose is to build a predictive model for diabetes classification. Different Predictive modeling techniques like logistic regression, k-nearest neighbor regression, decision tree classifier will be created to assess each model accurate prediction of diabetes. Python will be used for performing the modeling and analysis. RStudio can be used. Tableau or Power Bi for visualization and presentation of final recommendation.

Reference list:

[1] International Diabetes Federation. (n.d.). *Diabetes facts & figures*. Retrieved January 27, 2025, from https://idf.org/about-diabetes/diabetes-facts-figures/

[2] World Health Organization. (n.d.). *Diabetes*. Retrieved January 27, 2025, from https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Factors%20that%20contribute%20to%20developing,tests%20with%20a%20healthcare%20provider.

[3] Centers for Disease Control and Prevention. (2023). *National Health and Nutrition Examination Survey, 2021–2023: Data brief 516*. Retrieved January 27, 2025, from https://www.cdc.gov/nchs/products/databriefs/db516.htm#:~:text=The%20age%2Dadjusted%20prevalence%20of%20total%20diabetes%20increased%20from%209.7,in%20August%202021%E2%80%93August%202023