Replication 2025

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harm of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk. While there are different types of diabetes, type II diabetes is the most common form, and its prevalence varies by age, education, income, location, race, and other social determinants of health. Much of the burden of the disease falls on those of lower socioeconomic status as well. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approaching $400 billion dollars annually.

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. A clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables.

The **selected features** from the BRFSS 2015 dataset are:

High Blood Pressure

- Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional --> HighBP

High Cholesterol

- Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> HighChol
- Cholesterol check within past five years --> CholCheck

BMI

- Body Mass Index (BMI)

Smoking

- Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] --> Smoker

Other Chronic Health Conditions

- (Ever told) you had a stroke. --> Stroke
- Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) --> HeartDiseaseorAttack

Physical Activity

- Adults who reported doing physical activity or exercise during the past 30 days other than their regular job --> PhysActivity

Diet

- Consume Fruit 1 or more times per day --> Fruits
- Consume Vegetables 1 or more times per day --> Veggies

Alcohol Consumption

- Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) --> HvyAlcoholConsump

Health Care

- Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? --> AnyHealthcare
- Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? --> NoDocbcCost

Health General and Mental Health

- Would you say that in general your health is: --> GenHlth
- Now thinking about your mental health, which includes stress, depression, and problems with emotions, how many days during the past 30 days was your mental health not good? --> MentHlth
- Now thinking about your physical health, which includes physical illness and injury, how many days during the past 30 days was your physical health not good? --> PhysHlth
- Do you have serious difficulty walking or climbing stairs? --> DiffWalk

Demographics

- Indicate sex of respondent. --> Sex
- Fourteen-level age category --> _Age
- What is the highest grade or year of school you completed? --> Education
- Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") --> Income

Task

We plan to build a data center in Cork to store the above diabetes information. This data center has 5 nodes in PSA architecture. Create a replica set for this data center. This replica set should be named according to the location of this data center. To ease the whole task, keep all replica sets in the minimum recommended configuration.

1. Create this replica set manually, then take a screenshot of every step you perform in the process. You need to import your dataset into the primary node in the replica set.
2. With the replica set you have just created, explain how this replica set behaves in terms of automated failover and how to increase the size of each oplog to 25Gb.
3. How do you verify and globally reconfigure the write operation so that it is completed with two data-bearing members before returning? How do you reconfigure the read operation so that the replica set returning the most recent data when we perform a read operation?
4. How to support geographically local read operations in this replica set?
5. How do you remove this replica set for this data center?