Diabetes Cluster 2025

Introduction

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 (0 = no high blood pressure, 1 = high blood pressure)

High Cholesterol

- Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> HighChol (0 = no high cholesterol, 1 = high cholesterol)
- Cholesterol check within past five years --> CholCheck (0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years)

BMI

- Body Mass Index (BMI). The BMI ranges between 18 and 24, this is described as the “healthy range”; between 25 and 29 this is described as overweight; between 30 and 39 this is described as obesity; over 40 this describes as severe obesity.

Smoking

- Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] --> Smoker (0 = no, 1 = yes)

Other Chronic Health Conditions

- (Ever told) you had a stroke. --> Stroke (0 = no, 1 = yes)
- Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) --> HeartDiseaseorAttack (0 = no, 1 = yes)

Physical Activity

- Adults who reported doing physical activity or exercise during the past 30 days other than their regular job --> PhysActivity (0 = no, 1 = yes)

Diet

- Consume Fruit 1 or more times per day --> Fruits (0 = no, 1 = yes)
- Consume Vegetables 1 or more times per day --> Veggies (0 = no, 1 = yes)

Alcohol Consumption

- Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) --> HvyAlcoholConsump (0 = no, 1 = yes)

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 (0 = no, 1 = yes)
- Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? --> NoDocbcCost (0 = no, 1 = yes)

Health General and Mental Health

- Would you say that in general your health is: --> GenHlth (1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor)
- 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 (scale 1 – 30 days)
- 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 (scale 1 – 30 days)
- Do you have serious difficulty walking or climbing stairs? --> DiffWalk (0 = no, 1 = yes)

Demographics

- Indicate sex of respondent. --> Sex (0 is female and 1 is male)
- Fourteen-level age category --> _Age (1 = 18-24, 2 = 25–29, 3 = 30–34, 4 = 35–39, 5 = 40–44, 6 = 45–49, 7 = 50–54, 8 = 55–59, 9 = 60-64, 10 = 65-69, 11 = 70–74, 12 = 75-79, 13 = 80 or older)
- What is the highest grade or year of school you completed? --> Education (1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate))
- Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") --> Income (1 = less than $10,000 2 = less than $15,000 3 = less than $20,000 4 = less than $25,000 5 = less than $35,000 6 = less than $50,000 7 = less than 75,000 8 = $75,000 or more)

Part 1 (Self-Deployed MongoDB Cluster)

Our MongoDB cluster consists of 12 nodes (or independent servers), distributed over four small data centers across four different regions of USA (north, south, east, west). Each data centre represents a shard of the cluster, in which all nodes are organized as a replica set with the minimum configuration of a master-slave architecture. The configuration server has a group of three nodes {cfg0, cfg1, cfg2} organized as a replica set. The cluster is interfaced to the potential clients via lightweight processes. On one hand, these processes abstract and hide from clients the complexity of the cluster storing the data. The processes present the entire cluster as a single logical node, abstracting the actual physical distribution of the data among the shards. On the other hand, the processes also provide access to the metadata. The cluster is “simulated”, in the sense that it is built up on top of a single physical machine.

A script-based procedure is provided to allow you to build up the cluster on your own physical machine. To set up a distributed environment to store the diabetes application data in this self-deployed MongoDB cluster. Write bat (or bash) files and javascripts to set up this distributed cluster on a local machine.

Once the cluster is fully set up, we are ready to inject the diabetes application data and to distribute the data. A MongoDB database can be created on the cluster named diabetesAnalysisDB. The database can be sharded using a shard key. You need to be able to execute the scripts to set up the cluster and to demonstrate the whole process of distributing data. Furthermore, you also need to explain what is happening when the scripts are executing.

To answer the following queries with the built cluster:

1. List out all pregnant young women (age below 30) in the normal weight, who had full college qualification and annually earn between 50000 and 70000, have full healthcare coverage, and they are in a general good health condition and live in a healthy lifestyle (i.e. do not drink, smoke, and eat healthily)

2. How many obese male pensioners (age above 60), who are in a poor general health condition and have been suffering severe stress every day in the past 30 days due to their low annual income (below $20,000), therefore, they cannot afford a healthcare insurance and cannot visit a doctor due to the cost.
3. For each age group and different weight scales, list out all smokers who has diabetes and have never attended any schools and annually earn less than $35,000, they are in a poor health condition and have high blood pressure, high cholesterol, stroke, and some heart conditions.

Part 2 (Atlas Cluster)

Set up an Atlas cluster for the above Diabetes application. Connect your mongosh shell with the Atlas cluster and complete the following transactions (Note, each question should be completed in one individual transaction):

1. Randomly pick a male patient, who heavily consumes alcohol daily, to record the amount of alcohol consumed into his profile.
2. Randomly pick a female patient, who is in an excellent health condition and actively take part in physical exercises, to record the number of hours they do exercises daily. Then pick another male patient, who is in a fair good health condition, increase his BMI by 11.