NWI-IMC012: Bayesian Networks and Causal Inference

Assignment 1: Diabetes Interventions

Andrew Schroeder *s1111686*
Daan Kersten *s1053750*
Sven Meijboom *s1054374*
Radboud University *03-10-2024*

Problem Domain

Diabetes is one of the most significant chronic health diseases in the United States, with impacts on millions of Americans and representing a massive financial burden on the economy. Diabetes is caused by an inability to regulate sugar levels in the blood. More specifically, when food is consumed, digestion breaks the food down into simpler sugars which are released into the bloodstream. The sugars in the bloodstream then trigger the pancreas to release insulin, which enables cells to use the sugars in the bloodstream for energy. Diabetes disrupts this process by either causing insufficient insulin to be produced by the pancreas (Type 1) or by preventing effective utilisation of the insulin (Type 2).

Heart disease, vision loss, kidney disease, and a variety of other adverse health effects are associated with consistently high levels of blood sugar. Concerning Type 2 diabetes, there is no cure though its impacts can be reduced by making a variety of lifestyle changes including losing weight, eating healthy, and staying active. Early diagnosis is an effective tool against Type 2 diabetes as it can lead to lifestyle changes and more effective treatment. As a result both predictive and causal diabetes models are advantageous tools that can be used by public health officials to inform policy and interventions. In this research project, we will attempt to develop such a causal model.

Research Questions

We will use data related to lifestyle and diabetes in the United States which is collected by the Center for Disease Control (CDC) to answer the following questions:

1. What is the causal relationship between the education level of a person and their chance of developing Type II diabetes?
2. What variables are most significant in determining an individual's chance of developing diabetes?

Data

To answer our research questions, our project will make use of the CDC Diabetes Health Indicators dataset hosted on the UCI Machine Learning Repository. This dataset was created by the Center for Disease Control (CDC) to better understand the relationship between lifestyle and Type II diabetes in the US. Each row of the dataset consists of a single person participating in the study and there are no missing values. The dataset contains 21 features, 253680 samples, and is composed primarily of binary, numerical, and ordinal data types. Fortunately there is no categorical data (unordered categories), which will make data analysis a smoother process. Given that our second research question is concerned with those variables that have the largest affect on an individual's chance of developing diabetes, we will use all 21 features of the dataset - these include the following: Diabetes_binary, HighBP, HighChol, ColCHeck, BMI, Smoker, Stroke, HeartDiseaseOrAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthCare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, and Income.