Applying Causality Toolkit to Real-world Datasets

Arun Gaonkar
Mentor : Roger Dev

Introduction

Everything in this universe happens for a reason and every action has a reaction.
Analyzing causality can help in medical diagnostic analysis, time series analysis, and strategic planning
In real-world datasets, variables are inter-related, implying subtle correlations, which makes causal analysis difficult.

Causal Toolkits

1. HPCC_Causality
2. Because
   - Visualization bundle
   - Dependence & Independence tests
   - Causal Direction Tests

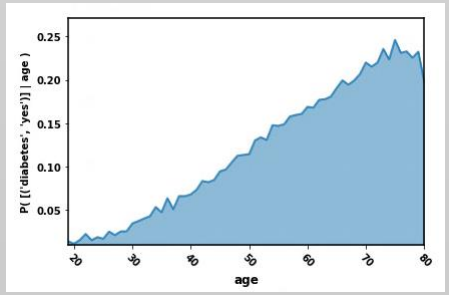
Analysis Steps

1. Finding & Analyzing Dataset
2. Pre-processing the dataset
3. Propose a Causal Hypothesis
4. Applying causal toolkits & analyzing
5. Interpretation & Causal Model
6. Hypothesis & Model Verification

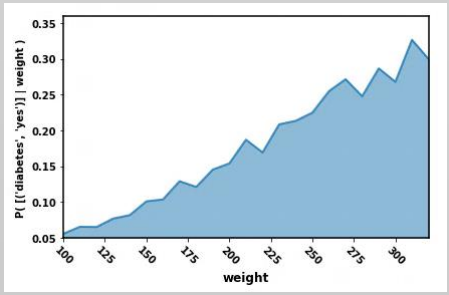
Causal Hypothesis Question

Proposed causal hypothesis question:
"What factors can influence the likelihood of a person having Diabetes?"

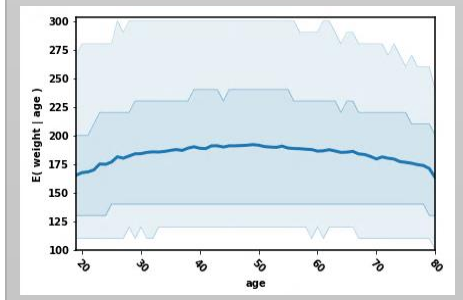
Analysis LLCP-CDC Dataset



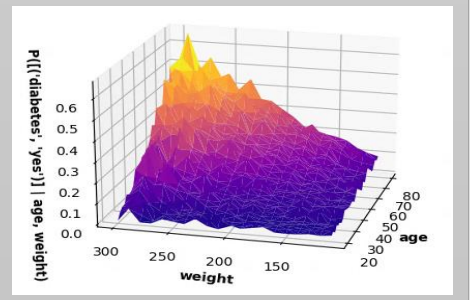
Diabetes vs Age



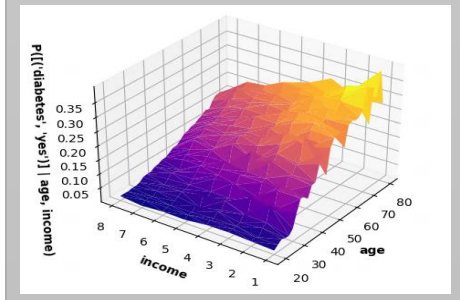
Diabetes vs Weight



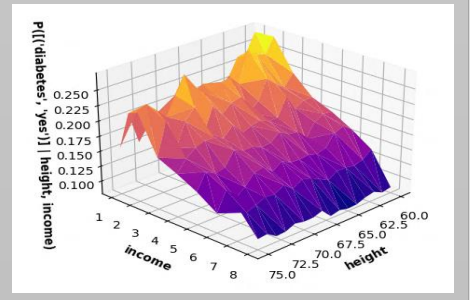
Weight vs Age



Diabetes vs Age, Weight

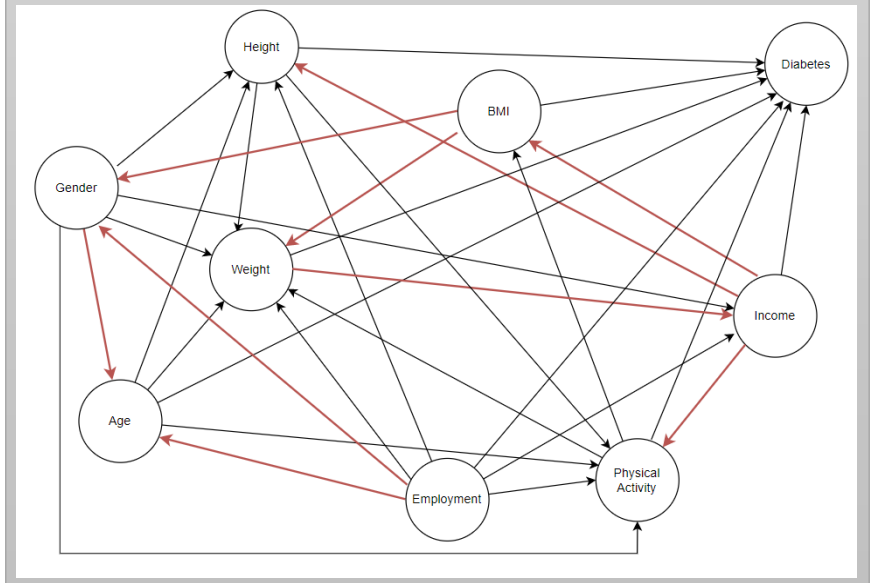


Diabetes vs Age, Income



Diabetes vs Height, Income

Causal Model



Conclusion

Factors like Age, Height, Weight, Income, type of Employment, Physical Activity, and gender have their effect on Diabetes.

- Most of the relations are practically and analytically correct.
- Some relations are unexpected, but probable valid proof can be generated.
- For some other relations, any explanation is rationally invalid.

Causality toolkit can be applied to analyze real-world datasets. But the cause-effect of latent variables cannot be incorporated into the causal model.