Intervention Methodology
Proposal


Radboud Universiteit

**Radboud Universiteit**

---

# DataMining Project Proposal

---

Groep 10

Dylan Elens, s1051438
Timo Walhof, s1037134

Oktober 2023

# Proposal

Diabetes is a global health concern with increasing prevalence, and its management and prevention are critical for improving public health. This project aims to analyze and leverage the Diabetes Health Indicators Dataset to gain insights into diabetes risk factors and understand the impact of lifestyle on diabetes. The primary objectives of this project are as follows:

1. Parse and understand the data
2. Visualize the data
3. Use Principle Component Analysis to find the most important predictors for Diabetes
4. Apriori algorithm to find causal relationships in the dataset

The dataset we want to use is the CDC Diabetes Health Indicators[1]. This is a really well-structured dataset, providing a comprehensive collection of healthcare statistics and lifestyle survey information related to diabetes. With 35 features encompassing demographics, lab test results, and survey responses for each individual, it offers a rich source of data for our analysis. Its well-organized format and accessibility make it an invaluable resource for understanding the relationships between various factors and diabetes diagnosis, ultimately contributing to our research in a highly effective manner. We wil also take inspiration from the paper[2] associated with this dataset

For this project, we intend to employ a range of essential tools and libraries in our data analysis and modeling processes. Python will serve as our primary programming language, offering versatility and a robust ecosystem for data science tasks. We'll use libraries like NumPy for numerical operations and Jupyter Notebooks for our coding environments. Additionally, we plan to leverage data visualization by using libraries such as Matplotlib, which will enable us to create informative graphs and charts to visually present our findings and insights throughout the project. These tools and libraries collectively form the foundation for our data-driven analysis.

# Bibliography

[1] Centers for Disease Control and Prevention (CDC). *CDC Diabetes Health Indicators Dataset*. https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. 2023.

[2] *Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico, 2000–2014*. https://www.cdc.gov/mmwr/volumes/66/wr/mm6643a2.htm. 2017.