**Public Health Data Visualization: Diabetes Risk Prediction Project**

**Use Case: Understanding the Stakeholders and Their Challenges**

Diabetes is a growing global health crisis, with millions at risk due to lifestyle choices and genetic predispositions. Early detection is crucial to prevent severe complications such as cardiovascular disease, kidney failure, and neuropathy. However, traditional screening methods are often inaccessible to large populations due to financial constraints and lack of awareness.

This project aimed to bridge the gap between raw medical data and actionable insights by developing a data visualization dashboard using R and Shiny. The objective was to assist multiple stakeholders in interpreting diabetes risk factors through interactive visualizations. The dataset was from the National Institute of Diabetes and Digestive and Kidney Diseases, specifically focusing on Pima Indian women aged 21 and older.

**Key Stakeholders and Their Challenges**

1. Healthcare Providers
   * Challenge: Difficulty in identifying high-risk individuals efficiently.
   * Need: A tool to quickly assess diabetes risk based on key health metrics.
   * Goal: Integrate visual analytics into patient screening and counseling.
2. Researchers
   * Challenge: Understanding the relationships between different risk factors.
   * Need: Statistical insights into how variables like glucose, BMI, and age contribute to diabetes risk.
   * Goal: Utilize the visualization platform to explore correlations and trends.
3. General Public (Patients and At-Risk Individuals)
   * Challenge: Lack of awareness about their personal diabetes risk.
   * Need: A user-friendly interface that allows self-assessment using personal health metrics.
   * Goal: Promote proactive health decisions and lifestyle adjustments.

**Data Cleaning and Preprocessing**

The dataset contained several inconsistencies, including missing values and zero-value entries in critical health metrics. Data cleaning was essential before meaningful insights could be extracted.

* Imported and explored the dataset using read\_csv() to understand its structure.
* I replaced zero values (Glucos, blood pressure, BMI, skin thickness, and insulin) with NA and removed those rows.
* Renamed columns for better readability.
* Checked for duplicates and removed any redundant records.

➡️ Outcome: A cleaned dataset with 392 observations ready for modeling and visualization. *(Detailed cleaning code is in my GitHub repository https://github.com/Shemelisyesuf/DiabetesPrediction)​.*

**Creating Static and Interactive Visualizations**

To effectively communicate the insights, a series of visualizations were designed:

1. Box Plot: Variable Distributions by Outcome
   * Helps in comparing the distribution of different health metrics between diabetic and non-diabetic individuals.
   * Key Finding: Diabetics generally have higher glucose and BMI levels. *(Visualization code in Data\_Visualisation.R​.)*
2. Correlation Heatmap
   * Displays the strength of relationships between numerical variables.
   * Key Finding: Glucose has the highest correlation (0.52) with diabetes, followed by BMI (0.35). Blood Pressure has a weak correlation.
   * Created using the corrplot package.
3. Pairwise Scatter Plots
   * Shows relationships between variables such as Age, BMI, and Glucose with diabetes risk.
   * Key Finding: Individuals with Glucose > 125 mg/dL and BMI > 30 were mostly diabetic.
4. Parallel Coordinates Plot (Interactive - Plotly)
   * Allows users to interactively explore multiple variables at once.
   * Key Finding: Clear trend of high glucose levels being a significant determinant.

Step 3: Building the Interactive Web Application (Shiny App)

The Shiny app was built to provide real-time interaction with the data. Users could:

✔ Adjust key health metrics using sliders and input fields.  
✔ View live predictions of diabetes risk.  
✔ Explore visual representations of data dynamically.

Implementation Details:

* Frontend: Used Shiny UI to create user input controls and real-time visualization panels.
* Backend: Implemented a Random Forest Model (randomForest package) to generate instant diabetes risk predictions.
* Visualization: Integrated ggplot2, plotly, and GGally for both static and interactive insights.

📌 *The complete implementation script can be found* <https://shemelis.shinyapps.io/My_project/>