Predicting Diabetes Risk Using R

Code

Shemelis Yesuf

2025-01-30

Introduction

Diabetes remains a global health challenge, and early detection is crucial for prevention and management. This project utilizes a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases to predict diabetes risk based on various medical indicators. The dataset consists of female patients (aged 21 or older) of Pima Indian heritage and includes variables such as glucose level, BMI, insulin, and pregnancy count. The objective is to develop a predictive model and an interactive Shiny app to help healthcare providers, researchers, and general users assess diabetes risk efficiently.

Understanding diabetes risk is critical for early detection and prevention, as diabetes is a leading cause of health complications worldwide. Data visualizations make it easier to:

1. Identify patterns and trends in health metrics, such as how glucose or BMI levels correlate with diabetes.
2. Communicate findings effectively to stakeholders, including healthcare professionals and patients.
3. Highlight key risk factors visually, aiding in education and decision-making.

Predictive models and visualiyation using shiny app are essential because they

- Enable early identification of at-risk individuals.
- Support personalized healthcare interventions.
- Empower individuals to understand their health metrics and make informed decisions.

Key User Persona, Aspired Actions, and Targets

The project is designed with three key user personas in mind:

1. Healthcare Providers

- **Aspired Actions**: Use the insights and tools to screen patients effectively and identify high-risk individuals.
- **Targets**: Focus on glucose, BMI, and age metrics for early interventions and personalized care.

2. Researchers

- **Aspired Actions**: Explore the relationships between risk factors and diabetes outcomes to develop predictive models.
- **Targets**: Understand the statistical significance and interaction of factors like pregnancies, age, glucose and risk of diabetes.

3. General Users (Patients)

- **Aspired Actions**: Use the interactive Shiny app to self-assess diabetes risk based on health metrics.
- **Targets**: Increase awareness of key risk factors and motivate lifestyle changes to reduce diabetes risk.

Significance:
This project aims to empower stakeholders by providing comprehensive, data-driven insights and practical tools that facilitate proactive diabetes management. By leveraging the power of data analytics and evidence-based approaches, the project seeks to enhance decision-making processes for healthcare providers, patients, and policymakers. These tools and insights will support early detection, effective intervention strategies, and personalized care plans, ultimately improving health outcomes and quality of life for individuals at risk of or living with diabetes. Additionally, the project aspires to foster a deeper understanding of diabetes trends and risk factors, enabling stakeholders to address the condition more effectively and promote preventive measures within communities.

Data Cleaning

The data cleaning process started by installing the necessary packages and loading essential libraries such as **tidyverse**, **readr**, and **dplyr**. These libraries are critical for data manipulation, visualization, and analysis.

Steps in Data Cleaning

- **Importing the Dataset**: The dataset was imported using the `read_csv` function from the `readr` package. Initial exploration involved viewing the structure and summary of the dataset to understand its dimensions and identify potential issues.
- **Zero Value Check**: Columns except for **Pregnancies** and **Outcome** were checked for invalid zero values. Zero values in columns such as **Glucose**, **BMI**, **Blood Pressure**, **Skin Thickness**, and **Insulin** were treated as missing data and removed.
- **Removing Invalid Rows**: Rows containing zero values in the aforementioned columns were filtered out to improve data quality.
- **Removing Duplicates**: Duplicate rows were identified and removed to avoid redundancy and bias in the analysis.
- **Renaming Columns**: Columns were renamed for better readability. For example, `BloodPressure` was renamed to `Blood Pressure`, and `SkinThickness` was renamed to `Skin Thickness`.

Hide

Hide

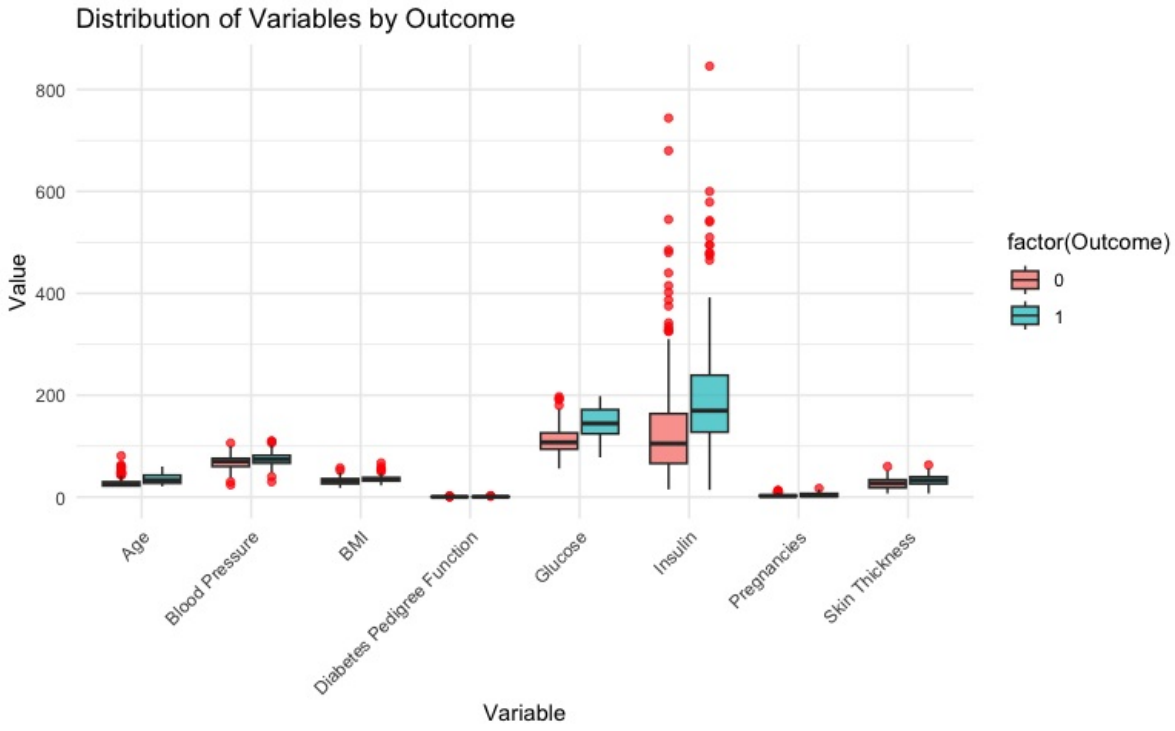
The cleaned dataset contains **392 observations** after removing invalid and duplicate entries.

Data Visualizations

Visualizations provide insights into the relationships between risk factors and diabetes outcomes.

Box Plot: Variable Distributions by Outcome

This plot illustrates how variables differ between diabetic and non-diabetic groups.



Interpretation:
This box plot provides a comparative view of the distribution of key variables for diabetic and non-diabetic individuals. For instance, the median glucose level for diabetic individuals is significantly higher, around **140 mg/dL**, compared to about **100 mg/dL** for non-diabetic individuals. Similarly, BMI values for diabetic individuals tend to be higher, with the interquartile range showing more overlap, highlighting that BMI alone may not be a definitive predictor. Blood pressure and insulin levels also show subtle but noticeable differences, whereas variables like skin thickness exhibit significant overlap, indicating weaker predictive power.

Correlation Heatmap

The correlation heatmap shows the strength of relationships between numerical variables.

Interpretation:
The heatmap reveals that glucose has the strongest correlation with diabetes outcomes (correlation coefficient: **0.52**), followed by BMI (**0.35**) and age (**0.26**). These numbers highlight glucose as a key predictor, whereas blood pressure and skin thickness exhibit weaker correlations (**< 0.1**) with diabetes. This insight reinforces the importance of glucose and BMI in prediction models and suggests a minor role for other variables.

Pairwise Scatter Plots

Scatter plots visualize pairwise relationships between key variables and the outcome.

Interpretation: The scatter plots reveal a clear separation between diabetic and non-diabetic individuals for glucose and BMI, where diabetic individuals are clustered in higher ranges. For example, individuals with glucose levels above **125 mg/dL** and BMI over **30 kg/m²** are predominantly diabetic. However, variables like age and pregnancies show broader overlaps, indicating their role as secondary predictors. These interactions between variables provide valuable insights into combined effects on diabetes risk.