Predicting Diabetes Outcomes Using Data Analytics and Machine Learning

William Kim

ALL-SPICE ALLIANCE
Supporting Pacific Impact through Computational Excellence

3 GOOD HEALTH AND WELL-BEING

Introduction

Diabetes is a major health concern, particularly among the Pima Indian population, who are genetically predisposed to the condition. People with diabetes have a higher risk of health problems including heart attack, stroke, and kidney failure (World Health Organization, 2023).Understanding the factors that contribute to diabetes development can improve early diagnosis and prevention. This research explores the relationship between medical and demographic variables and their predictive power in diabetes diagnosis.

Research Question

How do medical and demographic factors impact diabetes prediction, and which machine learning model provides the most accurate classification?

Methods

- Dataset: The Pima Indians Diabetes Dataset (Kaggle, originally from NIDDK), containing 768 samples with 9 variables.
- Predictor Variables: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.
- Target Variable: Diabetes diagnosis (1 = Diabetes, 0 = No Diabetes).

Results

Figure 1

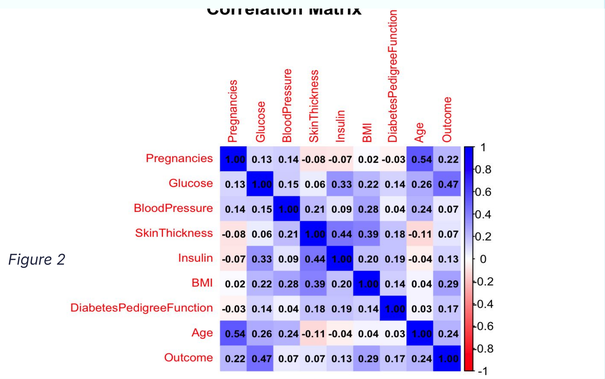


Figure 2

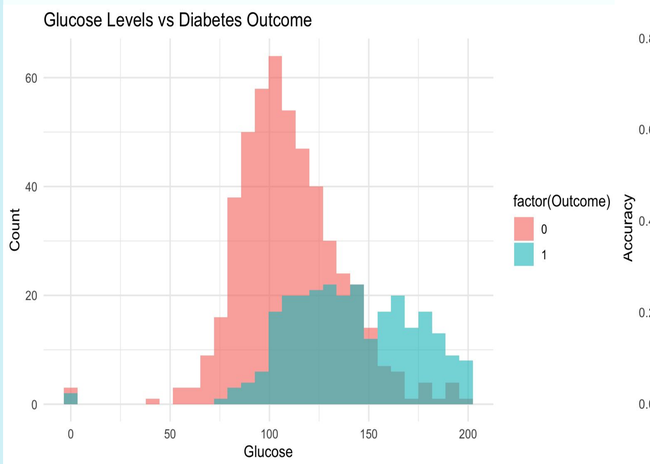
- Data Analysis & Visualization: Used R libraries (ggplot2, corrplot, dplyr) to explore variable distributions and correlations.
- Machine Learning Models: Logistic Regression, Random Forest, Neural Network, Naive Bayes.
- Shiny App: Developed an interactive tool for real-time diabetes prediction

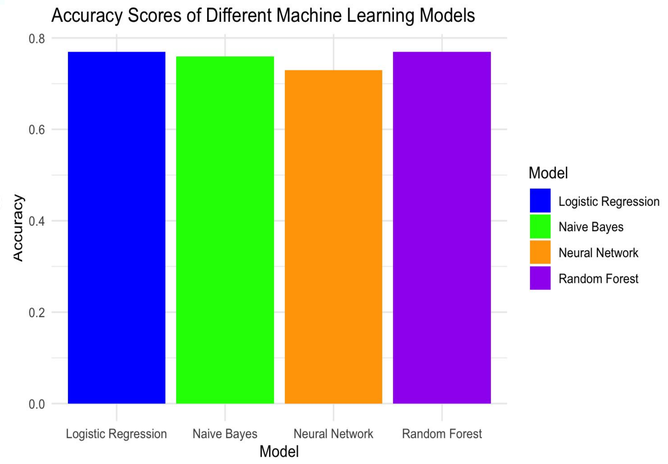
Key Predictors and Their Impact:

- Glucose Levels: Strongest predictor; high levels significantly increase diabetes risk.
- Blood Pressure, Skin Thickness, and Insulin: Showed weaker predictive power due to data inconsistencies.

Model Accuracy Scores:

- Logistic Regression: 77%
- Random Forest: 77%
- Neural Network: 73%
- Naive Bayes: 76%





Discussion & Future Work

This study demonstrates that glucose, BMI, and age are the strongest predictors of diabetes in the Pima Indian population. Logistic Regression and Random Forest models performed best in predicting diabetes outcomes. Future work includes refining data preprocessing, testing additional machine learning models, and improving the Shiny app for broader usability.

Acknowledgements

I want to thank the Data Science, Analytics, and Visualization program, Professor Mariah Yelenick, Dr. Amber Camp, and lastly Dr. Rylan Chong for supporting this project. This work was partially supported by the grant numbers HRD-2217242 (INCLUDES Alliance ALL SPICE) and PEARL DUE-2030654 (S-STEM). The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF.

Reference

- World Health Organization. (2023). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

Contact info

wkim151416@gmail.com
+1 (808) 206-1708
https://github.com/williamk670