Predicting the risk of diabetes

STATS/CSE 780 Course Project

Pao Zhu Vivian Hsu (400547994)
McMaster University

2023-11-30

Motivation

- Diabetes is a disease that occurs when the body cannot effectively produce or use insulin to regulate blood sugar levels.
  - 422 million people have diabetes worldwide and 1.5 million deaths that occur every year are linked to diabetes (World Health Organization 2023).
- Machine learning techniques are being used to predict diabetes.
  - Islam et al.'s study compares 3 different techniques and states that their decision tree produced the most accurate results (2020).
- *GOAL*:
  - Reproduce the decision tree in Islam et al.'s study to verify accuracy
  - Develop an SVM model and assess whether an SVM better predicts diabetes compared to a decision tree

Data

- Collected by Islam et al. from a hospital in Bangladesh (2020) and openly published on Kaggle (Larxel 2023).
- Data contains 17 variables and 520 observations
  - Response is binary and indicates whether the patient has a positive or negative risk for diabetes
  - Other variables describe the patient and presence of diabetes symptoms (ex. weakness, obesity)
- No missing data, no correlation between variables
- Outliers for age variable capped at 79 years using 1.5 IQR rule
- Imbalance in response variable

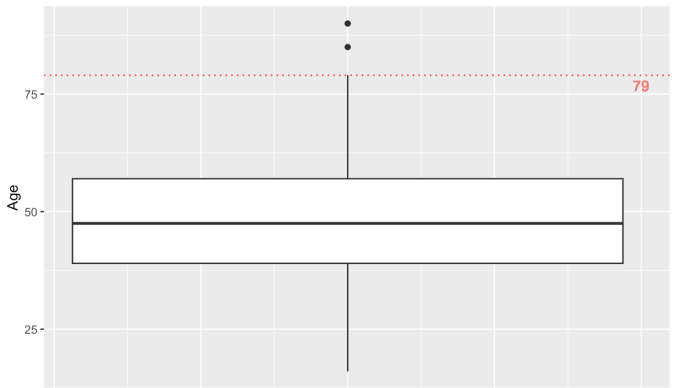

Figure 1: Box plot of age

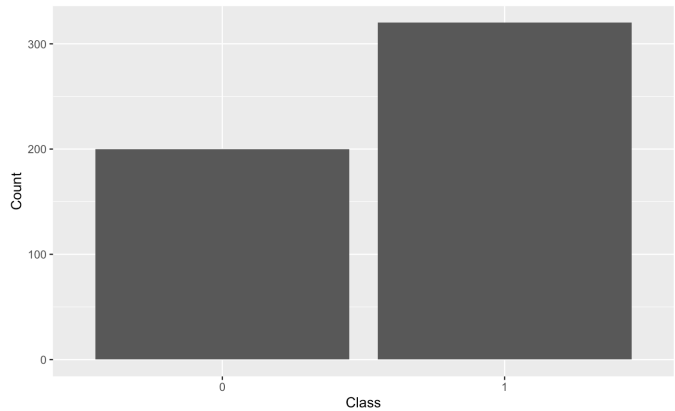

Figure 2: Bar chart of response

Methods (Decision Tree)

- Data was split in half for the training and testing sets
- First fit of the decision tree had a terminal node size of 16 with 94.62% accuracy
- Cross validation suggested a size of 12 terminal nodes instead
  - Accuracy remained the same
  - Pruning was still applied to reduce the cost complexity
- Random forest ensembling with m = 4 randomly sampled variables improved the accuracy to 98.85%

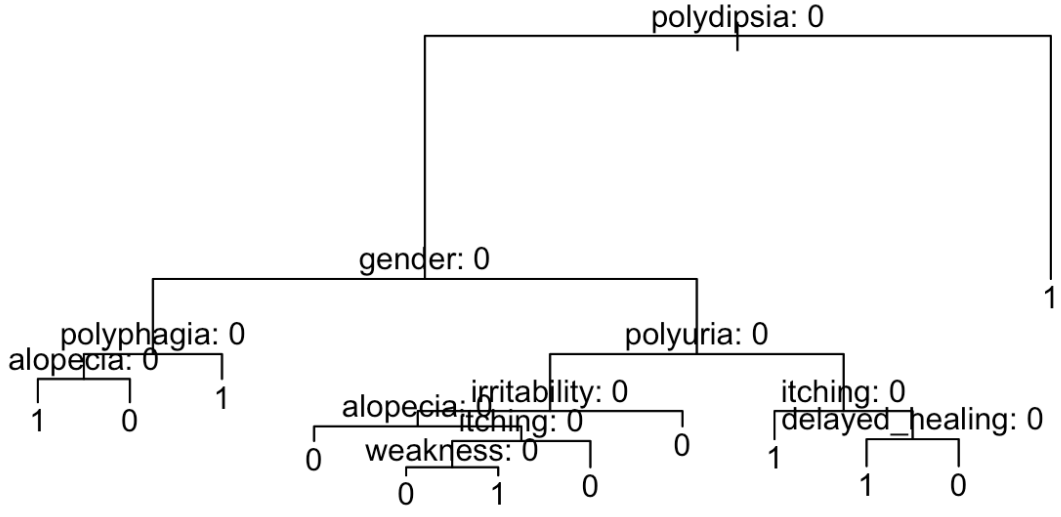

Figure 3: Decision tree after cross validation

Methods (SVM)

- The data was first scaled to ensure units are between 0 and 1 across all variables
- SVM was performed multiple times using kernel adjustment and cross validation for cost. Best models per kernel:
  - Linear - cost = 3, accuracy = 93.85%
  - Polynomial: cost = 150, accuracy = 95.77%
  - Radial basis function (RBF): cost = 50, accuracy = 97.69%

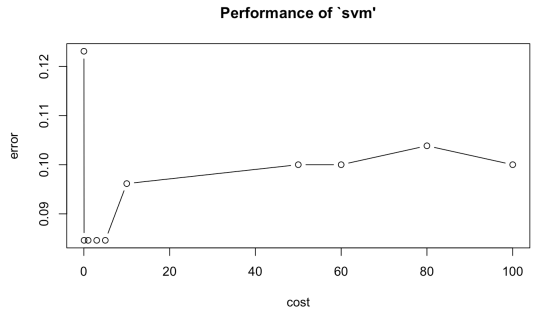

Figure 4: CV for linear SVMs

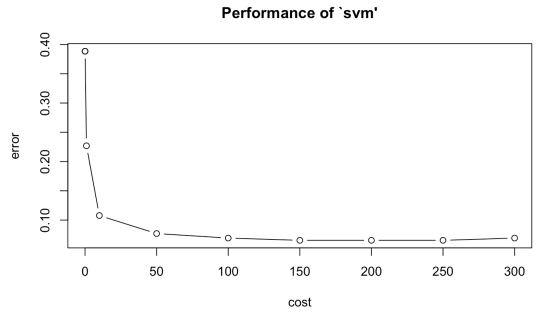

Figure 5: CV for polynomial SVMs

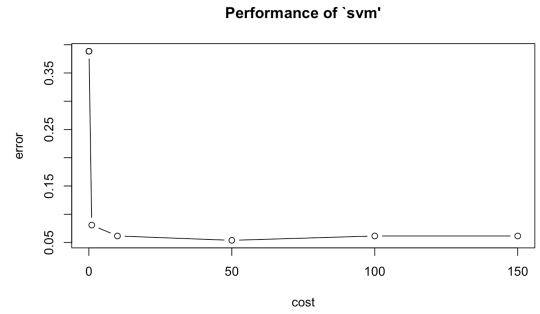

Figure 6: CV for RBF SVMs

Results

- Both methods had high accuracy
- Random forest outperformed SVM
- In terms of Islam et. al's study (2020), decision trees remain the best method to predict diabetes compared to SVM and 2 other methods
- Patients who are male, have poluria, and have polydispia are at a greater risk for diabetes

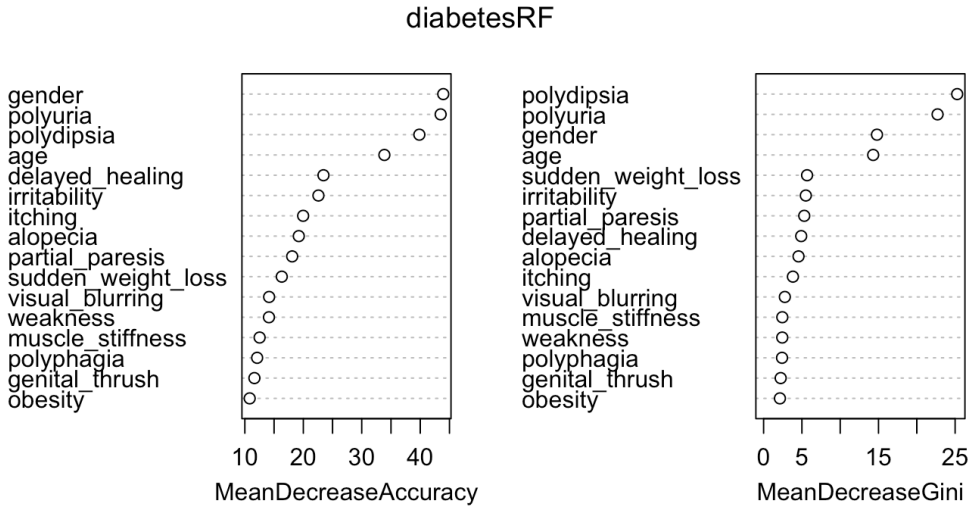

Figure 7: Importance of variables from random forest

Discussion

- Random forest
  - Stability of the results may be improved by increasing the number of trees
- SVM model
  - Selecting the best cost is a balance between bias and variance.
    - The linear model had a cost of 3 -> high bias and low variance
    - The polynomial model had a cost of 150 -> low bias and high variance
    - Both may not be a great fit despite having high accuracy; RBF is more balanced
    - Repeated cross-validation may give help us choose better costs
  - Selecting the best kernel can be a challenge for SVM models due to the risk of overfitting
    - Further investigation on the pattern of the data may be useful to improve kernel tuning

Thank You!

References

Islam, M. M. Faniqul, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. 2020. "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques." In *Computer Vision and Machine Intelligence in Medical Image Analysis*, edited by Mousumi Gupta, Debanjan Konar, Siddhartha Bhattacharyya, and Sambhunath Biswas, 113–25. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12.

Larxel. 2023. "Early Classification of Diabetes." https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

World Health Organization. 2023. "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1.