Comparing Diabetes Risk Assessment Scores

Creating predictive systems for T2DM risk by machine learning, weighting factors and classical pen and paper:

Authors: Hanyuan Zhang, Zhihao Zhang, Ka Fan, Joshua Tsai

Abstract

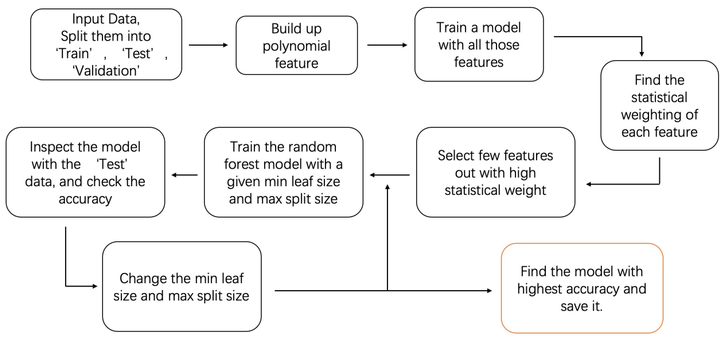
We created three models to predict diabetes outcomes based on a public data set of n=768 Pima Indian females. The models were based on a tally score, a weighted factor model, and a random forest model and had accuracies of 66.2%, 76.6%, and 76.6%, respectively. The sensitivity of each model was 0.7, 0.52, and 0.63, respectively. The specificity of each model was 0.64, 0.9, and 0.84. Overall, each model had its own advantages, but since we intended this as a screening tool, the pen and paper model appears most accurate upon initial survey, which will be useful for clinicians to quickly calculate. However, closer inspection reveals a likely tradeoff between sensitivity, so the factor weighting/machine learning models gave probabilistic outputs that could be modified to increase sensitivity in exchange for specificity. 35% of the population had diabetes and 65% did not have diabetes.

Introduction

Diabetes Mellitus Type 2 (T2DM) is a chronic metabolic disease involving resistance to the hormone insulin that afflicts nearly 3.8 million in the UK and is one of the most significant comorbidities for a range of diseases including cardiovascular disease, strokes, and obesity. Although it currently has no cure, several factors including weight loss are known to prevent or even lead to remission of the disease, implying that if we can identify target patients, there is potential to focus efforts on prophylactic treatments. T2DM currently costs the NHS billions to treat, meaning there is a great need for these sorts of efforts.

Methodology

- Random forest is the advanced version of a binary classification decision tree that builds many slightly different decision trees and combines them like a forest, where each tree is weighted by its accuracy. To make our model more accurate, we found *polynomial features* of our eight arguments which multiplied them to form a second-degree polynomial. However, we only found second order features to avoid overfitting.



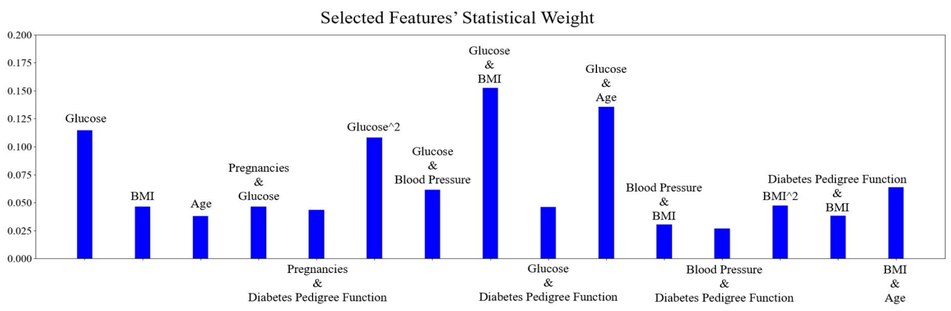


Fig.1. The statistic weight of each feature in the random forest model.

- The factor weighting method used optimized parameters guided by clinical literature research and the random forest models to determine factor weightings. Probabilistic predictions were rounded with a bias factor to become binary predictions for diabetes.
- Based on medical literature, we chose certain parameters for our pen and paper method. We averaged the weightings given to each parameter in existing medical literature to develop an average weighting for our model (rounded to a whole number for ease of calculation). Scores were added up and an optimized cut-off for diabetes prediction determined.
- Methods were compared by tallying how many outcome predictions were correct and then comparing how many true/false negatives/positives each method provided. Data was also stratified to better describe it.

Results

- The models were based on a tally score, a weighted factor model, and a random forest model and had accuracies of 66.2%, 77.9%, and 79.2%, respectively.
- Sensitivities were 0.7, 0.52, and 0.593, respectively. Specificities were 0.64, 0.9, and 0.9.
- These results compare decently to the medical literature ADA(79,67), AUDRISK(74,67), and FINDRISC(78,77). Accuracy appeared to be better than random for all tests. For a verification set composed of the high blood glucose population, accuracy fell in the factor-weighting model to 73.3%. Examining the new subgroups with the pen/paper method showed a change of sensitivity and specificity of 72%,57% for the glucose subgroup and 58%,74% for the BMI subgroup.

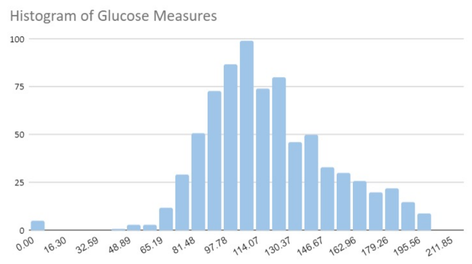


Fig.2. Distribution of glucose measures amongst population

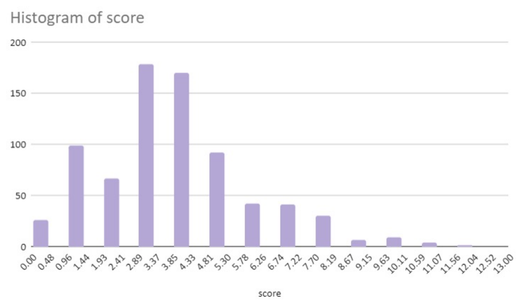


Fig.3. Score by classical (pen and paper) method of population

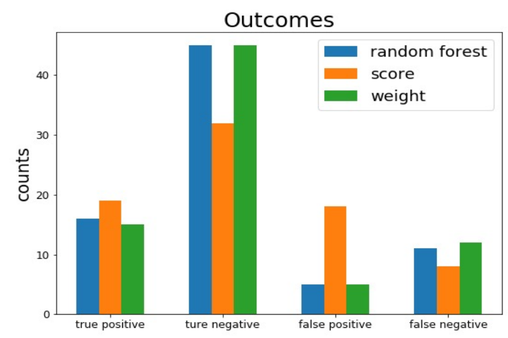


Fig.4. Comparison between three models.

Conclusions

- While difficult to use in a triad setting, the random forest model was the most accurate at describing data. This is not surprising as it has the most adjustable parameters.
- However, the factor weighting model came in at a close second, with a 90% specificity, which could make it useful for differentiating between diabetes and a similar illness. Of course, by further modifying the specificity/sensitivity curve to convert our probabilities into binary predictions, we could have improved these methods. Unsurprisingly, when using subset populations, accuracy decreased since our models were not built for this. Unfortunately, it appeared that accuracy decreased more in populations with risk factors than those who don't have risk factors, implying our models might be biased by the proportion of diabetes in the sample.
- Surprisingly, the pen and paper method, while least accurate, was most sensitive and best as a screening tool.

References

1. Waugh N, Shyangdan D, Taylor-Phillips S, Suri G, Hall B. Cambridge Risk Score. 2013 [cited 2021 Mar 20]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK260920/
2. Pima Indians Diabetes Database | Kaggle [Internet]. [cited 2021 Mar 20]. Available from: https://www.kaggle.com/uciml/pima-indians-diabetes-database
3. The Australian Type 2 Diabetes Risk Assessment Tool (AUSDRISK) | Australian Government Department of Health [Internet]. [cited 2021 Mar 20]. Available from: https://www.health.gov.au/resources/apps-and-tools/the-australian-type-2-diabetes-risk-assessment-tool-ausdrisk
4. Zhang L, Zhang Z, Zhang Y, Hu G, Chen L. Evaluation of Finnish diabetes risk score in screening undiagnosed diabetes and prediabetes among U.S. adults by gender and race: NHANES 1999-2010. PLoS One. 2014 May 22.

Interdisciplinary Research Computing Course – Group project poster
Group Number: Group 6

Project repository: https://github.com/tian-xuan/Interdisplinary-computing-projrct