6 - Final Project Phase 2 High Fidelity Prototype

Ente Kang, Jingwen Ji

Link to GitHub: https://github.com/entekang/6-FinalProject

Introduction

Diabetes is a chronic condition that affects millions of people, and the number of adults diagnosed with diabetes has more than doubled over the past 20 years [1]. Currently, there is no cure, however, some known risk factors include age and being overweight.

Thus, our team hopes to answer the following questions: (1) How can repeated measurements of biomarkers (blood pressure, BMI, cholesterol levels etc.) help with the classification of diabetes status. (2) What are the risk factors that influence time-to-diabetes.

(1) How can repeated measurements of biomarkers help classify diabetes status?

Numerous patients will have repeated assessments of biomarkers that can help predict diabetes status. If a biological test is retaken, there usually is a good reason for it. The test results and the number of tests contains rich information on the patients' medical conditions. For example, if a doctor was worried about a patient's health and their risk of developing a condition, multiple tests would be ordered. We seek to include both the number of repeated biomarker tests and the change in their test results to assist with the classification of the status of diabetes. From this, we will be able to answer the following: (1) The effect of a unit change in the biological test on the odds of developing diabetes. (2) Do more tests signal an increased or decreased risk of diabetes? (3) The association between change in age on the development of diabetes. The answers to these questions will help us gain a better understanding of the diagnosis of diabetes by utilizing longitudinal information from a patient.

(2) What are the risk factors that influence time-to-diabetes?

Given a study period, some patients will have developed diabetes, while others have not. For those who have not, it does not mean that they cannot have diabetes in the future, it is just that we will not observe it during the study time-period. However, we can gain valuable information on the time it takes for a patient to develop diabetes and the associated risk factors. This is inherently different from a classification task, as we would be able to know when the patient may experience diabetes. Knowing these risk factors is invaluable for clinicians, especially if it is modifiable. For example, if we were to determine that a high BMI was associated with a shorter time-to-diabetes, then the doctor could recommend a change in lifestyle habits for the patient, which would then lead to improved health outcomes.

Methods

We will be using the Diabetes 10K dataset, which has a sample size of 10,000 observations. Within these 10,000 observations, there are repeated measurements for a patient, effectively giving us a sample 1,099 patients. The number of repeated measurements range from 1 to 8, so some will have more biological tests taken than others. There is also information on one's age at the time, and the time that some of these biological tests were taken.

We calculated the change in the predictor variable by subtracting the first measurement from the last measurement of everyone for each of age, sBP, BMI, LDL, HDL, A1C, TB, FBS and total cholesterol. In other words, the change in age is the age at the last measurement minus the age in the first measurement. We felt it was more important to look at the change in predictors, because this could tell us how an increase or decrease in measurements over time might affect the development of diabetes. Even though the diabetes status did not change over time for patients in the study, the repeated predictor measurements still need to be taken into account. The change in a predictor provides more information than using the mean or median or simply taking a random sample of a patients' biological tests as it gives a sense of the patient's progression. Out of 1099 observations, there were missing values for LDL (10), HDL (5), TG (3), and Total Cholesterol (19). We imputed using the median, as the median is less prone to outliers than the mean. We then used visualizations to explore which of these predictors may be associated with the presence of diabetes. Ones that looked promising were then used as predictors in our classification model.

Since we are interested in the statistical effects of our predictor variables on the classification of diabetes, Logistic Regression [7,9] was used. It is worth noting that there are many other classification models such as KNN, and Decision Trees, but these will not allow us to see the effect of our predictors on the outcome

Results

The predictors for our final model include the changes in: age at examination, A1c, FBS, Total Cholesterol, and the number of repeated biological tests.

At the 0.05 level, a one unit increase in the change in age at examination was associated with an 1.1 times increase in odds of developing diabetes, holding everything else constant. In other words, individuals whose age increased by 3 years vs. 2 years at the end of the study were more likely to develop diabetes. This was expected, as age can be a big contributing factor to diabetes [1].

In terms of pure predictions, we saw that the performance of Logistic Regression was mediocre at best, and improvements can be made in terms of model performance [2]. Since it can be very devastating to tell patients who have diabetes that they don't have the condition, we focused on the recall metric. The recall was 100% but the overall precision was only 72% on the test data. Logistic regression was not able to correctly predict any true negatives. In other words, our model said that patients who truly did not have diabetes had diabetes, which is not ideal. We also implemented decision trees [4] to compare its performance with logistic regression and saw that it was able to predict more true negatives. We decided on the decision tree because it is non-parametric, meaning that there are no assumptions about the distribution of the data and overall structure of the model, which is quite a contrast vs. logistic regression. Furthermore, it can handle missing values, and has the ease of interpretation. The decision tree correctly predicted 46 true negatives, and 169 true positives given a test set of 330 patients. The recall was 70% and a precision of 79%. Improvements are needed for this model as well.

Discussion

At this phase, we have built two classifiers (Logistic Regression and Decision Tree) to predict diabetes status using the changes in: age at examination, A1c, FBS, Total Cholesterol, and the number of repeated biological tests. The purpose of choosing Logistic Regression was to be able to see the predictor variables' effects on the odds of developing diabetes. The Decision was then implemented to compare predictive performance against logistic regression.

We saw that Logistic regression is not suitable here, as it was not able to correctly predict any true negatives. On the other hand, the Decision Tree performed a bit better in the sense that it was able to correctly predict some true negatives and true positives.

Our whole purpose of using the change in predictor variables was to combat the repeated measurements of patients – which inherently contains a lot of rich sequential information. A big limitation for us is that at this point, we do not have the necessary tools to deal with repeated measurements on the response and predictor variables from a machine learning perspective. We plan to address this limitation as soon as we are given the necessary tools to combat repeated longitudinal measurements.

As for our next steps, we hope to make improvements on our diabetes classifier and transform the Diabetes 10K dataset for time-to-event analysis and prognostic modeling. It is hypothesized that ensemble methods may yield better prediction results than

Logistic Regression [5,6,8[. We will also be looking at the time it takes to develop diabetes through the Cox model [13] and AFT model [14], after adjusting for relevant risk factors. Our motivation is that even if patients do not have diabetes right now, it does not mean that they cannot develop it later in life. It is just that we may not observe it during the study period, thus leading to censoring. With survival analysis techniques, we can assess either the hazards of the event occurring or the acceleration of the time-to-event, which will be very helpful from a clinical perspective. Some preliminary exploration [12] would include the Kaplan-Meier curves, model fitting, and model diagnostics. If the concordance is greater than 0.8 [11], then prognostic modeling will be considered.

Individual Contributions

Ente Kang: Data processing, modeling, coding, report writing

Jingwen Ji: Comprehensive literature review and summary of information

References

1. https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=With%20diabetes%2C%20your%20body%20either,releases%20it%20into%20your%20bloodstream.
2. Lai H et al. Predictive Models for Diabetes Mellitus using Machine Learning Techniques. *BMC Endocrine Disorders* 2019:19
3. Garcia JF et al. Noninvasive Prototype for Type 2 Diabetes Detection. *Journal of Healthcare Engineering* 2021
4. Fregoso-Aparicio L et al. Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Systematic Review. *Diabetology and Metabolic Syndrome* 2021;13
5. Deberneh H, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International journal of Environmental Research and Public Health* 2021;18
6. Rejendra P, Latifi S. Prediction of Diabetes using Logistic Regression and Ensemble Techniques. *Computer Methods and Programs in Biomedicine Update* 2021
7. Abdollahi J, Nouri-Moghaddam B. Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction. *Iran Journal of Computer Science* 2022;5;205-220
8. Daghistani T, Alshammari R. Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes. *Journal of Advances in Information Technology* 2020;11
9. Zabor E. et al. Logistic Regression in Clinical Studies. *Int J Radiation Oncol Biol Phys* 2022;112;271-277
10. Ambrish G. et al. Logistic Regression Technique for Prediction of Cardiovascular Disease. *Global Transitions Proceedings* 2022;127-130
11. Harrell FE et al. Evaluating the Yield of Medical Tests. *American Medical Association* 1982;18;2543-2546
12. Clark TG et al. Survival Analysis Part I: Basic Concepts and First Analyses. *Br J Cancer* 2003;89;232-238
13. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society* 1972;34;187-220
14. Wei LJ, The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis. *Statistics in Medicine* 1992