Heinz 95-845:
Using Machine Learning to Predict Undiagnosed Diabetes in the NHANES Dataset

Kevin McGrady   KEVINMCG@ANDREW.CMU.EDU

Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Abstract

The abstract is the summary of the article. Your potential readers will glance at the abstract to decide if the article is worth reading. Make it good–this is your most-read text!

1. Introduction

The American Diabetes Association states that 29 million Americans have diabetes and of that, 8 million are undiagnosed. Diabetes ranks seventh in leading causes of death in the United States [1]. Detecting individuals that are likely to have undiagnosed diabetes and properly diagnosing them will help decrease the risk of diabetes-related complications like cardiovascular disease [2]. Application of machine learning algorithms can help determine which individuals should be screened.

Yu et al. used support vector machines to classify individuals into diabetes and non-diabetes groupings [3]. Omogbai used support vector machines with canonical correlation analysis [4]. Dall et al. used logistic regression [5]. Jain used a feed forward neural network [6]. Barber et al. cite 18 models in their review of pre-diabetes risk assessment tools - 11 used logistic regression, 6 used decision trees, and 1 used support vector machines (Yu) [7].

In contrast, this analysis applies multiple machine learning methods to execute a classification task that is strictly focused on the prediction of undiagnosed diabetes. Yu and Omogbai sought to assign the correct label between (A) diagnosed or undiagnosed diabetes versus pre-diabetes or no diabetes and (B) undiagnosed diabetes or pre-diabetes versus no diabetes. Dall had a multi-class prediction, assigning no diabetes, pre-diabetes, and undiagnosed diabetes. Jain estimated individual level-risk of diabetes.

In Section 2, Background, information is provided on diabetes, machine learning, and NHANES. In Section 3, Model, information is provided on the selected machine learning algorithms.

The layout of the paper is as follows: Background (Section 2), Model (Section 3), Methods (Section 4), Results (Section 5), Discussion (Section 6), and
Conclusion (Section 7).

2. Background

Diabetes is the common name for the metabolic disease, diabetes mellitus. The disease arises from problems within the relationship between the pancreas, insulin, glucose (sugar), the bloodstream, and the cells within the bloodstream. An improper balance of blood sugar can have many deleterious effects. There are three primary forms of diabetes: Type 1, Type

2, and gestational [8-10]. For this analysis, there is no differentiation made between Type 1 and Type 2 diabetes. Gestational diabetes is not included.

Machine learning, in this application, entails solving a classification task. Computer algorithms are employed to predict whether to classify an individual as an undiagnosed diabetic or not. For this study, six different algorithms are applied. More information on the types of algorithms used in this study can be found in Section 3, Model.

NHANES is the National Health and Nutrition Examination Survey issued by the National Center for Health Statistics, which is part of the Centers for Disease Control and Prevention [11]. The NHANES components selected for this study were questionnaire, laboratory, demographics, and examination. For more information on NHANES data subset used in this analysis, refer to Sections 4-4.3.

3. Model

This analysis applies logistic regression, decision tree, naive bayes, tree augmented naive bayes, support vector machines, and random forest algorithms to the task of predicting which individuals have undiagnosed diabetes. The entirety of the code is written in the R Project for Statistical Computing. The libraries used are as follows: stats, rpart, bnlearn, e1071, and randomForest. Each algorithm was used "as is" or "off the shelf"; no hyperparameter tuning was conducted. Neural network algorithms were not applied to this dataset as that level of complexity did not seem appropriate for the relatively small dataset and the R implementation of neural networks is suboptimal. Documentation for these libraries used in this study can be found at `https://cran.r-project.org/web/packages`. All code for this project is available at `https://github.com/95845k/project_rnhanes_dm`.

4. Methods

The following subsections explain how the data was extracted (4.1), how the primary outcome was derived (4.2), how features were selected and transformed (4.3), and how the models were evaluated (4.4).

4.1 Data Extraction

Data for this project was extracted from NHANES using the R library RNHANES [12]. Selected files were downloaded from the period 1999-2000 through 2013-2014 (see: `https://github.com/95845k/project_rnhanes_dm/blob/master/prd_code_data_extract`.R. Survey response codes were re-coded to response descriptions and checked (see: `https://github.com/95845k/project_rnhanes_dm/blob/master/prd_code_data_extract_checks`.R. After joining all of the datasets, 82,091 records were present.

4.2 Outcome Derivation

Following Yu, the study dataset was filtered individuals aged 20 and above who were not pregnant at the time of the survey [3]. The age filter eliminated 38,298 records and following that, the pregnancy filter removed 1,416.

The primary outcome was determined by both questionnaire and laboratory results. If the response was anything other than "No" to the question, "Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes?" then the record was removed. This question posed as the proxy for diabetes diagnosis. The study dataset included only those individuals that did not have a diabetes diagnosis. The result of this question removed 5,757 records. To determine whether an individual had diabetes three laboratory tests were employed: fasting plasma glucose (FPG), A1c, and oral glucose tolerance test (OGTT). Following Dall, diabetes was defined as having either FPG greater than or equal 126 or A1c greater than or equal to 6.5 or OGTT greater than or equal to 200 [5]. If an individual did not have any of the three lab values present, then that individual record was eliminated from the dataset. This removed 3,527 records. Those that answered "No" to diagnosis and had a lab value in those ranges were considered to be undiagnosed diabetics.

The final study dataset contained 33,093 records. The population therein is non-pregnant individuals age 20 or over in the NHANES dataset between 1999 and 2014, who have not been diagnosed with diabetes, and who had diabetes laboratory test results. The outcome is binary: undiagnosed diabetic or not. (see: `https://github.com/95845k/project_rnhanes_dm/blob/master/prd_code_data_transform.R`

4.3 Feature Choices

Model features were based on a review of the literature [3-7, 13-14]. The features chosen to predict which individuals had undiagnosed diabetes were age, gender, race/ethnicity, education level, family history of diabetes, body mass index, waist circumference, high blood pressure, high cholesterol, average number of alcoholic drinks consumed, smoking status, and number of hours of sleep. Age, body mass index, waist circumference, average number of alcoholic drinks consumed, and number of hours of sleep were employed as real values. Race/ethnicity was transformed into four categories: hispanic, white, black, and other. Education was transformed into three levels: no high school degree, high school degree, and college degree. High school degree includes those individuals with associates degrees. Family history of diabetes, high blood pressure, high cholesterol, and smoking status were treated as binary. For smoking, responses of "Every day" or "Some days" were coded as true values.

In addition to the upfront missing values, response descriptions of "Refused" and "Don't Know" were treated as missing. There was no evidence to suggest that the upfront missing values of any feature were missing not at random (MNAR). Additional features addressing missingness were deemed to be non-essential for this study. If the number of "Refused" and "Don't Know" responses had been more substantial, there perhaps would have been more of a case for MNAR. As it stands, when excluding "Don't Know" for family history, these responses made up less than 1 percent of the data. The imputation method for values missing at random used in this study was multivariate imputation by chained equations. The mice package in R provided the algorithm. Three imputations and three iterations were the settings. Results were pooled. The most frequent value was used for non-real values. In case of a tie, the first value was taken. The mean imputed value was used for

real values. (see: `https://github.com/95845k/project_rnhanes_dm/blob/master/prd_code_data_transform.R`

For the naive bayes and tree augmented naive bayes algorithms, every features are required to be discrete. A "discretize" function was applied to the dataset. This transformed the real valued features into discrete buckets.

4.4 Evaluation Criteria

In order to evaluate the performance of the selected machine learning algorithms on this classification task, the study dataset was split into training and testing datasets. Seventy percent of the data was used for training (n = 23,165). Thirty percent was reserved for future testing (n = 9,928). The training set was then further divided into a base training dataset and a validation dataset. Seventy percent of the training data was used a on base training set (n = 16,215). Thirty percent was used for validation (n = 6,950). The first goal of evaluation is have the models perform well on the validation dataset. The best performing model would then be selected and tested using the testing dataset.

Because of class imbalance, only five percent of the individual records are represented as undiagnosed diabetes, accuracy is a poor measure of evaluation. If a model predicts "no diabetes" every time, the model will be correct approximately 95 percent of the time. The true positive rate or recall, measuring how many correct predictions of undiagnosed diabetes were made divided by the the actual number of undiagnosed diabetes records, is a better measure. In addition, the false positive rate, and precision or positive predictive value will be evaluated. Precision is the number of correct predictions of undiagnosed diabetes divided by the number of undiagnosed diabetes predictions made. These measures will be illustrated in confusion matrices, receiver operator characteristic (ROC) curves, and precision-recall curves. For ROC curves, the area under the curve (AUC) will be evaluated.

5. Results

Present the results here. Do not describe how the results were obtained. Those descriptions belong in Section **??**.

Typically there are multiple parts and subparts of your study. Use subsections to report the results.

Give us some numbers about how well your method works, especially in comparison to some baselines. You should provide a summary of the results in the text, as well as in figures (such as figure 1).

6. Discussion and Related Work

This is where you characterize the outcomes of your method and draw conclusions from you experiment. The discussion will build upon the Introduction and the Results sections to synthesize where your contribution brings the field. Discuss any implications of your work. Discuss limitations of your work. Are there situations where you should and should not use your method. What implications are there on policy making, clinical decision making, or future research activities? Remember to contextualize your work with respect to related work and provide references.

Figure 1: Example smile graphic.

7. Conclusion

Summarize your work one more time, this time assuming the reader has read your paper. Build suspense for what your next extension to this method would be.

8. References

1. http://www.diabetes.org/diabetes-basics/statistics
2. American Diabetes Association. "Standards of medical care in diabetes - 2013." Diabetes care 36. Supplement 1 (2013): S13.
3. Yu, Wei, et al. "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes." BMC Medical Informatics and Decision Making 10.1 (2010): 16.
4. Omogbai, Aileme. "Application of multiview techniques to NHANES dataset." arXiv preprint arXiv:1608.04783 (2016).
5. Dall, Timothy M., et al. "Detecting type 2 diabetes and prediabetes among asymptomatic adults in the United States: modeling American Diabetes Association versus US Preventive Services Task Force diabetes screening guidelines." Population health metrics 12.1 (2014): 12.
6. Jain, Deepti, and Divakar Singh. "A Neural Network based Approach for the Diabetes Risk Estimation." International Journal of Computer Applications 73.10 (2013).
7. Barber, Shaun R., et al. "Risk assessment tools for detecting those with pre-diabetes: a systematic review." Diabetes research and clinical practice 105.1 (2014): 1-13.
8. https://en.wikipedia.org/wiki/Diabetes_mellitus
9. http://www.webmd.com/diabetes/guide/diabetes-basics
10. https://my.clevelandclinic.org/health/articles/diabetes-mellitus-an-overview
11. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
12. https://cran.r-project.org/web/packages/RNHANES/vignettes/introduction.html
13. Stiglic, Gregor, and Majda Pajnkihar. "Evaluation of major online diabetes risk calculators and computerized predictive models." PloS one 10.11 (2015): e0142827.
14. Knutson, Kristen L., and Eve Van Cauter. "Associations between sleep loss and increased risk of obesity and diabetes." Annals of the New York Academy of Sciences 1129.1 (2008): 287-304.

Appendix A.

Some more details about those methods, so we can actually reproduce them.