Heinz 95-845: Using Machine Learning to Predict Undiagnosed Diabetes in Individuals 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.

The paper layout is as follows: Background (Section 2), Model (Section 3), Methods (Section 4), Outcome Selection (Section 4.1), Data Extraction (Section 4.2), Feature Choices (Section 4.3), Comparison Methods (Section 4.4), Evaluation Criteria (Section 4.5), 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]. For more information on the NHANES dataset, refer to the Sections 4-4.3.

3. Your Model Name Here

This section describes your model and references the notation you introduced in the Background Section. **Figures are definitely helpful here**, so that someone who is in your area can visualize how your approach is novel, and someone who is not in your area can visualize what you are doing.

If you introduce new mathematical or statistical methods, use the terminology you defined in Section **??** and define your model. Give the technical details and remember: do be precise and do be concise.

If you are combining existing methods, then you don't need to provide a ton of detail: feel free to just cite other packages and papers and tell us how you put them together.

If you developed new code that does not (should not) contain sensitive or private information, include a reference, e.g.:

" Code is available at `http://my.github.page.com` "

4. Experimental Setup

Note: if the paper is more about the application than the method, this Section may be entitled Methods and appear before Section **??**

By reading the Experimental Setup, your reader should have the information necessary to replicate the study.

Describe the cohort/data. Provide information about the population, the inclusion and exclusion criteria, what data were extracted, how features were processed, etc. In fact, you may want the following headings. **A flow chart can be very helpful** to illustrate the experimental setup, study design, inclusion/exclusion process, etc.

For more clinical application papers, each of the sections above might be several paragraphs or pages because we really want to understand the setting.

4.1 Cohort Selection

Describe how the samples you used were selected to form your cohort and also to provide cohort descriptive statistics. In methodologic papers, the "Table 1" describing the population by covariate summary statistics goes here. In application papers, "Table 1" leads the Results Section. Relevant information about the study design, such how cases and controls were identified, goes here. See Section **??** for an example of how to build a table in LaTeX.

4.2 Data Extraction

Describe the pipeline from raw data to processed data. Figures can be helpful. What assumptions did you make? How did you deal with missing data? Do not place interpretations here except possibly for short justification phrases. Longer discussions about the assumption you made go in the Discussion Section.

4.3 Feature Choices

What features were used? What conversions were necessary? What assumptions (e.g. i.i.d.) are made? with how you might have converted the raw data into features that were used in your algorithm.

4.4 Comparison Methods

To evaluate your model, often times you will compare against existing models. If so, include them here with a brief description, citation, and any tweaks you made for your experiment.

4.5 Evaluation Criteria

Evaluation methods belong here as well. Perhaps you used accuracy and the AUROC– explain why these are most useful measures of the outcome.

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.

5.1 Results on Application A

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 tables (such as table **??**) and figures (such as figure **??**).

You may use subfigures/wrapfigures (LaTeX packages) so that figures don't have to span the whole page or multiple figures are side by side.

Method	Outcome (%)
Us	20.1
Baseline	18.2

Table 1: Outcome by method used. These are our results.

5.2 Results on Application B

Did more than one experiment type?

Figure 1: Example smile graphic.

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.

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. Bibliography

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

Appendix A.

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