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

Discovering individuals who have undiagnosed diabetes is a pursuit designed to improve health outcomes. By combining the National Health and Nutrition Examination Survey (NHANES) and a suite of machine learning algorithms this goal can be achieved. This paper produces results that show competitive performance and offers avenues for future enhancement.

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. Section 4, Methods, describes the data extraction, outcome derivation, feature selection, and evaluation processes. Section 5 provides the results and Section 6 brings forth a discussion and opportunities for future work. Section 6 is the conclusion.

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.

Prior to any statistics or application of machine learning, Table 1 provides some clues as to what separates the "no diabetes" group from the "undiagnosed diabetes group." Older, male, Hispanic non-smokers with larger waist sizes, no high school degree, a family history of diabetes, high blood pressure, and high cholesterol appear to be more likely to be undiagnosed diabetics.

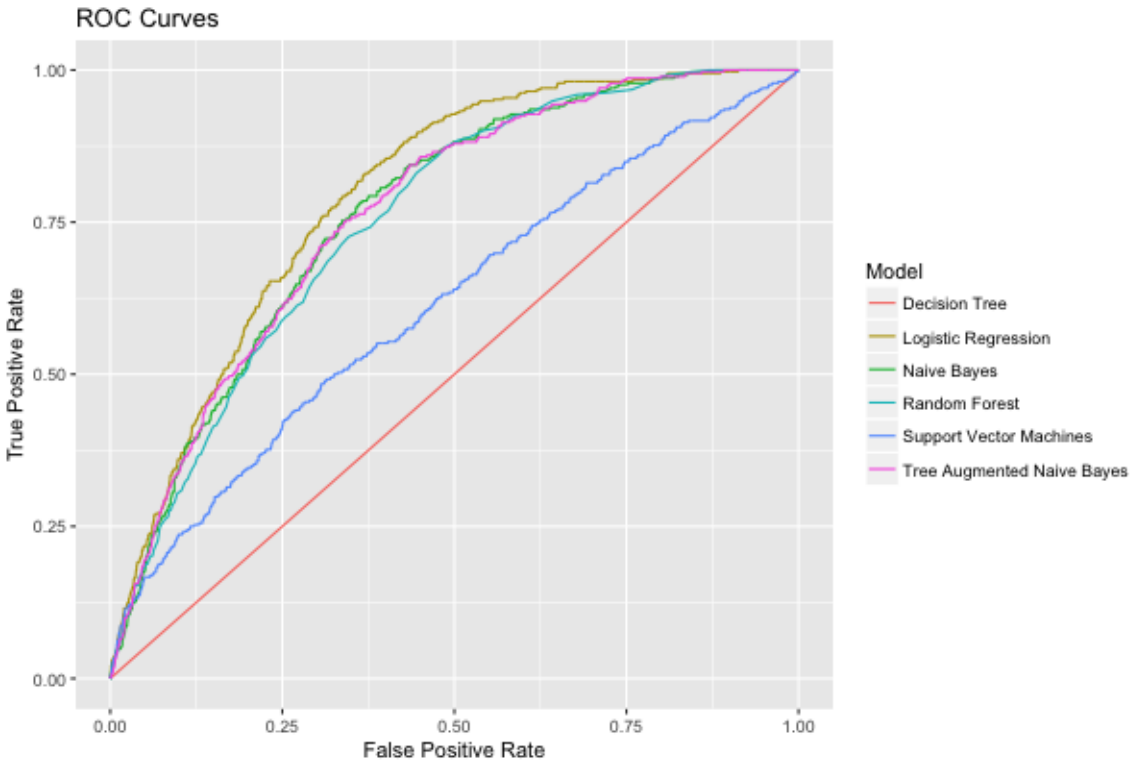
Less than five percent (775/16,215 in the base training set) of the individual records are represented as undiagnosed diabetes. Because of this class imbalance, accuracy is a poor measure of evaluation. If a model predicts "no diabetes" every time, the model will be correct approximately 95 percent of 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 receiver operator characteristic (ROC) curves, precision-recall curves, and confusion matrices. For ROC curves, the area under the curve (AUC) will be evaluated. Both the ROC and the precision-recall curves calculate their respective metrics over a range of probability thresholds. The confusion matrix displays the prediction and outcome relationship at one probability cutoff. A probability threshold of fifty percent was used for the confusion matrices in this paper. The model has to return a probability greater than fifty percent in order to classify an individual as an undiagnosed diabetic. Anything less than that, the individual will be classified as having no diabetes.

Table 1: Descriptive Statistics on Base Training Dataset

		Outcome	
	Category	No Diabetes	Undiagnosed Diabetes
n		15440	775
Gender (%)	Male	7649 (49.5)	422 (54.5)
	Female	7791 (50.5)	353 (45.5)
Race/Ethnicity (%)	White	7585 (49.1)	335 (43.2)
	Hispanic	3795 (24.6)	225 (29.0)
	Black	3018 (19.5)	162 (20.9)
	Other	1042 (6.7)	53 (6.8)
Education (%)	HS Degree	8013 (51.9)	369 (47.6)
	No HS Degree	4040 (26.2)	313 (40.4)
	College Degree	3387 (21.9)	93 (12.0)
Family History of Diabetes (%)	No	9402 (60.9)	407 (52.5)
	Yes	6038 (39.1)	368 (47.5)
High Blood Pressure (%)	No	11008 (71.3)	355 (45.8)
	Yes	4432 (28.7)	420 (54.2)
High Cholesterol (%)	No	10738 (69.5)	430 (55.5)
	Yes	4702 (30.5)	345 (44.5)
Smoker (%)	No	7917 (51.3)	533 (68.8)
	Yes	7523 (48.7)	242 (31.2)
Age (mean (sd))		47.80 (18.05)	60.88 (14.85)
BMI (mean (sd))		28.09 (6.37)	31.86 (7.38)
Waist Size (mean (sd))		96.41 (15.26)	107.76 (15.57)
Number of Alcoholic Drinks (mean (sd))		2.82 (2.58)	2.63 (2.24)
Hours of Sleep (mean (sd))		6.84 (1.23)	6.88 (1.44)

5. Results

The results on the training validation set are displayed below. The ROC diagram shows that logistic regression performed the best. The AUC for the logistic regression model was 0.794, which tops the 0.765 AUC of naive bayes and tree augmented naive bayes.



Moving from left to right, the precision-recall curves shows that at most levels of recall, logistic regression is the most precise (is on top).

The confusion matrix for logistic regression tells a different story (Table 2). Of the 6,950 records in the training validation set, 372 were undiagnosed diabetes and 6,578 were no diabetes. At the fifty percent threshold, the logistic regression model did not return any predictions of undiagnosed diabetes. The class imbalance was not overcome, and so the model predicted "no diabetes" each and every time for an accuracy of ninety-five percent. This is true for all models with the exception of naive bayes. Naive bayes predicted undiagnosed diabetes 164 times, 28 were true positives and 136 were false positives (Table 3). The true positive rate for logistic regression is 0. The true positive rate for naive bayes is eight percent. The false positive rate for logistic regression is 0. The false positive rate for naive bayes is two percent.

Based on the evidence, logistic regression performed the best. However, as shown in the confusion matrix, the results are rather unsatisfying. Naive bayes offers an alternative, but the true positive rate is not high enough to warrant its selection. Moreover, logistic regression provides more interpretable results.

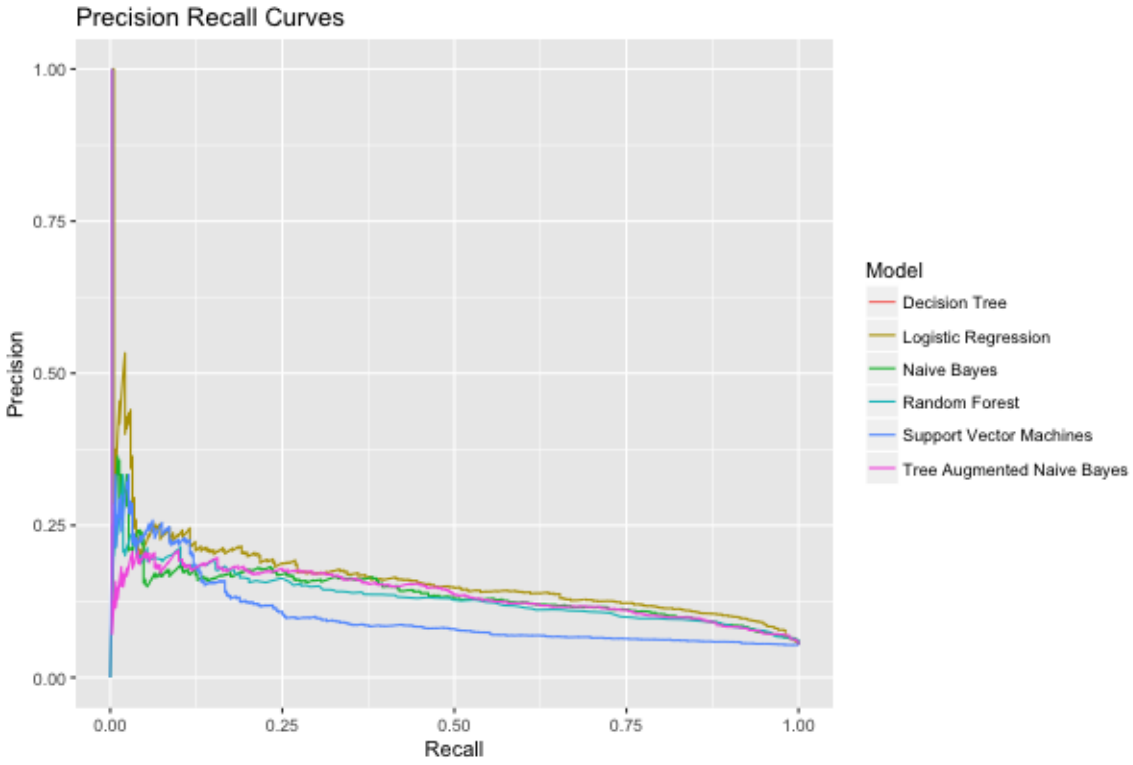


Table 2: Logistic Regression Confusion Matrix

		Actual Outcome	
		Undiagnosed Diabetes	No Diabetes
Predicted Outcome	Undiagnosed Diabetes	0	0
	No Diabetes	372	6578

Table 3: Naïve Bayes Confusion Matrix

		Actual Outcome	
		Undiagnosed Diabetes	No Diabetes
Predicted Outcome	Undiagnosed Diabetes	28	136
	No Diabetes	344	6442

Table 4: Logistic Regression Results

Variable	Exp. Estimate	P-value	
(Intercept)	0.00004	0.00000	***
Gender:Female	0.91223	0.30289	
Race/Ethnicity:Black	1.47521	0.00032	***
Race/Ethnicity:Hispanic	1.75804	0.00000	***
Race/Ethnicity:Other	2.56819	0.00000	***
Education:College Degree	0.67106	0.00117	**
Education:No HS Degree	1.29292	0.00380	**
Family History of Diabetes:Yes	1.43778	0.00000	***
High Blood Pressure:Yes	1.33910	0.00055	***
High Cholesterol:Yes	1.05581	0.49828	
Smoker:Yes	0.98595	0.88120	
Age	1.04185	0.00000	***
BMI	0.99323	0.59292	
Waist Size	1.04331	0.00000	***
Number of Alcoholic Drinks	1.00086	0.96286	
Hours of Sleep	1.01826	0.54317	

Much of what appeared in Table 1 has been validated by the results from the logistic regression model (Table 4). Minority races and ethnicities have higher odds of being in the undiagnosed diabetes group. Hispanics are 1.76 times more likely to be undiagnosed diabetics than whites, holding non-race/ethnicity factors constant (p-value = 0.000). Individuals without a high school degree are 1.29 times more likely (p-value = 0.003) to be undiagnosed diabetics than high school graduates, while college graduates are 0.67 times less likely. A family history of diabetes and having high blood pressure also increase the odds, respectively.

6. Discussion and Related Work

The AUC for the logistic regression model appears to be reasonable. Yu published an AUC of 0.732 for the classification scheme 2, which incorporated both undiagnosed diabetes and prediabetes [3]. The maximum AUC shown reported in Barber for models using logistic regression for pre-diabetes was 0.744 [7]. The significant factors and odds ratios also appear to be reasonable. For undiagnosed diabetes and total prediabetes, Dall reports a 1.97 odds ratio for Hispanics (compared to whites) and a 1.36 odds ratio for having high blood pressure. The contribution of this study is in the focus on the undiagnosed diabetes population, the application of multiple machine learning algorithms, and the missing value imputation (which seems absent in all referenced studies).

However, the positive predictive value offered by the logistic regression model at a threshold probability of fifty percent is valueless. The true positive rate is zero. In order to achieve a true positive rate of seventy-percent, the threshold probability has to be reduced to under five percent. This also comes at the expense of thirty percent false positive rate. Therefore, I will keep the remaining test set "locked away."

Further analysis is required before final results can be published. The class imbalance problem must be resolved. Undersampling or oversampling or synthetic minority oversample (SMOTE) are approaches designed to overcome class imbalance [15]. For logistic regression, there is also opportunity to add interaction terms and quadratic variants. Feature reduction, through principal or independent component analysis is an option as are other feature engineering techniques. The missing value imputations could be checked by running the models on a complete case dataset. Following that, there are opportunities for optimizing and hyperparameter tuning of the algorithms as opposed to the "off the shelf" approach used in this study. More models could be explored and applied including boosting algorithms and additional ensemble models. An alternative approach could be implemented - changing the task from classification to regression. Lab values could be scaled and combined for the study dataset and features could be awarded predicted diabetes risk points through linear regression or regression trees.

7. Conclusion

This paper has produced contributions to the study and prediction of undiagnosed diabetes. There is further work to be conducted, which will yet further advance the pursuit of improving outcomes at a population health scale.

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.
15. https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/node6.html