Journal of High School Science

Developing a predictive model for the risk of diabetes

Ye Y[1]

Submitted: January 20, 2022, revised: version 1, March 1, 2022, version 2, April 27, 2022
Accepted: May 14, 2022

Abstract

The proportion of the population diagnosed with type II diabetes has gradually increased in the United States. It is necessary to find methods to better diagnose type II diabetes in the population. To this end, this report constructed two classification models, a logistic regression model and an artificial neural network, to predict the probability of type II diabetes, based on factors derived from family composition, family income and health conditions. The data in this study were taken from the responses of 31997 individuals to the National Health Interview Survey (NHIS). To ensure that the data was usable by the classification algorithms, mean value imputation and min-max feature scaling was applied to fill the missing values and re-scale the range of features in the data-cleaning process. The percentage of the data split into the training and test sets was 70 : 30 respectively. The two predictive models were further validated by an overall evaluation of the model, statistical tests of individual predictors, and an assessment of the relative importance of independent variables. Both, the artificial neural network and the logistic regression models exhibited similar sensitivity and selectivity, with the AUC scores from their Receptor-Operator Curves being 0.85 and 0.84 respectively. Hypertension and physical health had the strongest positive and the strongest negative correlation respectively with the diagnosis of diabetes. Consistent with Garson's algorithm, hypertension was the key variable predisposing to diabetes. Since hypertension is often primary, leading a healthier lifestyle may be a method to prevent type II diabetes. Exercising to balance energy intake and expenditure may be an effective way to decrease the probability of becoming diabetic.

Keywords

Risk of diabetes, Logistic regression model, Artificial neural network, Diabetes prediction, Hypertension, Receptor operator curve, Classification algorithm, Biomarkers, BMI, Cholesterol

[1]Corresponding author: Yutong Ye, Shanghai Starriver bilingual school, No. 2588, Jindu road, Minhang district, Shanghai, China. Iris20050217@163.com

Introduction

In recent years, diabetes, which refers to a chronic disease with symptoms of elevated blood sugar, becomes a severe problem in the United States. In 2000, the population that had been diagnosed with diabetes was 11 million (4% of the U.S. population), and recently the population drastically amounts to 34.2 million, which is 10.5% of the U.S. population (1). Over time, diabetes may lead to multiple organ failures including heart attacks and damage to nerves, eyes, feet, as well as kidneys (2). The public often believes that the greatest risk factor of diabetes is obesity (3). However, as shown by researchers, family income is one of the dominant factors related to the cause of diabetes.

Bird et.al. (4) investigated the relationship between socioeconomic status and type II diabetes along with the prevalence of type II diabetes in Korea. Since individuals' socioeconomic status affects their diets and lifestyles, they concluded that socioeconomic status is strongly related to type II diabetes. Using the Korean population as the base, Yelena tries to find the relation of diabetes diagnosis to individuals' demographic information. In addition, since there was no adequate information on type II diabetes in Canada, Hwang investigated the adjusted effects of individuals' income in leading to type II diabetes (5). Furthermore, they also discuss the long-term effects of type II diabetes to remind the public to contain the development of type II diabetes in an early stage.

The purpose of this study is to find the variables that most correlated with type II diabetes and to develop a model that helps researchers to predict the propensity for an individual will be diagnosed with diabetes in the future. The previous two groups of researchers have gaps that have not been discussed. This report will fill up those gaps. The study by Bird et.al. involved the Korean population, and this report includes different races as one of the variables, and so it is more likely to explain the most related variable to diabetes. Hwang's study was more concerned with the effects of diabetes rather than on the factors related to diabetes. This report focuses more on those factors, which may help researchers and healthcare professionals to make better diagnostic predictions.

In addition, we acknowledge that we are not the first group of people applying machine learning models to diagnosing diabetes. Adler et.al., for example, tested several machine learning models built to predict the presence of diabetes, based on a wide selection of biomarkers like blood sugar and gene markers (6). Even though they achieved relatively good predictive accuracy, the use of biomarkers in prediction models is invasive, time and labor consuming, involves doctors' visits and needs to clear regulatory and ethical hurdles. Therefore, this report tries to improve his research by only employing non-invasive physiological markers and obtaining a relatively high model performance at the same time.

Methods

Data

This report uses data from the National Health Interview Survey (NHIS) in 2019, which is a population-based survey established by the CDC to monitor the prevalence of the population health condition in the United States and to evaluate the effects of illness on individuals (7). The whole survey mainly encapsulates family composition, tobacco use, alcohol, and drug usage, as well as other dietary behaviors and physical activity questions. The data is collected by a face-to-face survey to random adults in households through asking demographic and other related questions. The 2019 NHIS dataset is used in this report. Before the data-cleaning process, the NHIS dataset has 31997 valid observations.

Item Code	Category	Question	Response
INCGRP_A	Family Income	Sample adult family income (grouped)	1= $0 to $34,999, 2= $35,000 to $49,999 3= $50,000 to $74,999, 4= $75,000 to $99,999 5=$100,000 Or Greater
SEX_A	Household Composition	Sex of Sample Adult	1= Male, 2=Female
AGEP_A	Household Composition	Age of Sample Adult	18-84= 18-84 Years 85+= 85+ Years
HISPALLP_A	Household Composition	Single and multiple race groups with Hispanic origin	1= Hispanic, 2= NH White, 3= NH Black, 4= NH Asian, 5-6= NH AIAN 7= Other Single/ Multiple Races
EDUC_A	Household Composition	Educational Level of Sample Adult	0= Never Attended, 1= Grade 1-11 2= 12 Grade, 3= GED, 4= High School Graduate, 5= Some College, No Degree 6-7= Associate Degree, 8= Bachelor's Degree 9= Master's Degree, 10= Professional School Degree, 11= Doctor Degree
SOCERRNDS_A	Social Functioning	Difficulty Doing Errands Alone	1= No Difficulty, 2= Some Difficulty 3= A Lot of Difficulty, 4= Cannot Do At All
HYPEV_A	Health Status	Ever Had Hypertension	1= Yes, 2=No
CHLEV_A	Health Status	Ever Had High Cholesterol	1= Yes, 2=No
PHSTAT_A	Health Status	General Health Status	1= Excellent, 2= Very Good, 3= Good 4= Fair, 5= Poor
BMICAT_A	Weight	Categorical Body Mass Index	1= Underweight, 2= Healthy Weight 3= Overweight, 4= Obese
WEIGHTLBTC_A	Weight	Weight Without Shoes	100-299 = 100-299 pounds
HEIGHTTC_A	Height	Total Height in Inches	59-76 = 59-76 inches
DIBEV_A	Diabetes	Ever Had Diabetes	1= Yes, 2=No

Table 1: Variables used in this study

Statistical Models

Figure 1 below shows the percentage of missing data of each variable in the dataset. The mean value imputation method (see Pre-processing below) was used to populate missing data points.

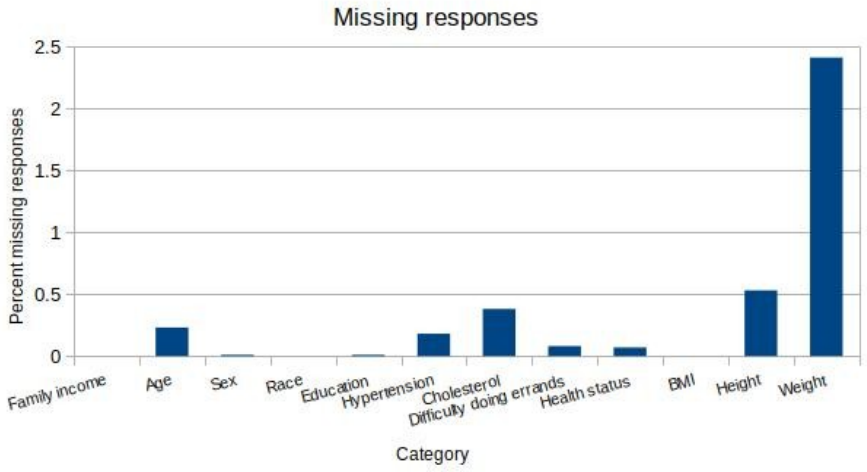

Figure 1: Percentage of missing values for each feature

Pre-processing

The data set was pre-processed in this step to improve both the training speed and accuracy. Since there is inevitably missing data, imputation was required to better analyze and extrapolate the missing data. As most machine learning algorithms are not able to deal with missing values, the missing values were replaced with the mean value of the entire feature column (mean value imputation). Some machine learning algorithms, such as artificial neural networks, require a specific technique called feature scaling which transforms different features into comparable scales for better training speed and accuracy. In this study, the min-max scalar was used for this purpose. For each feature, its minimum and maximum value are first computed as X_{min} and X_{max}. Then each data point X with respect to that feature is replaced by X_{sc} calculated as:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Using this equation, X_{sc} was the value on which analysis was performed.

Logistic Regression Model

The logistic regression model was first introduced by Joseph Berkson and he coined the term "logit" by analogy to the "probit" (8). A logistic regression model refers to a model that is used to predict the probability of an incidence to happen. The probability varies from 0 to 1, with zero indicating not likely to happen and one indicating very likely to happen. Instead of a linear relationship, the logistic regression model fits an "S" shape which can be expressed using the equation below:

$$\ln\left(\frac{y}{y-1}\right)=a_0+a_1x_1+a_2x_2+\cdots+a_nx_n$$

In the above equation, a_0 is the intercept, x_n represents the independent variables, and a_1 to a_n, are their corresponding coefficients (weights). In this study, the objective was to find the coefficients ($a_0 \ldots a_n$) minimizing the sum of squared errors (SSE) so that the predicted values deviate the least from the real values.

The L2 regularization was used to minimize the probability of model over-fitting. Over-fitting occurs when the predictive model has high accuracy on the training set but low accuracy on the test set. Regularization is a technique to control over-fitting by adding an additional term to the loss function such that the model coefficients do not take extreme values. L2 regularization adds an L2 penalty, which is equal to the square of the magnitude of coefficients, and the new loss function is represented in the expression below:

$$\min_{w,c}\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\log\left(\exp\left(-y_i\left(X_i^Tw+c\right)\right)+1\right)$$

In the above expression, w is the weights of all the variables. This study applied the L2 regularization to the Logistic regression model to minimize the deviation from actual data and the differences between the training set and test set.

Artificial neural networks

Artificial neural network was first invented in 1958 by psychologist Frank Rosenblatt, and it was intended to model human brain processes of visual data (9). An artificial neural network is a system that shows the interrelation of each variable (input) to the result (output) through layers and nodes. The system is inspired by the biological neural networks that the inputs will travel through the hidden layers when signaled and eventually send the information to the output layer. Its pattern-matching and learning capability allow researchers to address more complex problems. Unlike in biological systems, here the "signal" is a real number, and the output of each neuron is computed as the

sum of some non-linear functions applied on the inputs.

A typical artificial neural network consists of one input layer, several hidden layers, and one output layer. The input layer is the first layer, the output layer is the last layer, and any layers between them are hidden layers. The data are passed into the input layer, processed by the hidden layers, and finally transformed into predicted labels in the output layer. In this study, the model incorporated one hidden layer with three nodes.

In each layer, there are also edges connecting the nodes from the previous layer to the nodes in the current layer, and those edges are often used as weights during the calculation process. Like in logistic regression models, the goal for training artificial neural networks is to find a set of edges (weights) that minimize the cost function and to achieve the best prediction performance.

A package called "neuralnet" in R was used to conduct neural network analysis (10). The package neuralnet focuses on multi-layer perceptron, which is applicable when modeling functional relationships.

Model Validation

Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and a false negative (FN) is when the prediction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (11). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the

diagonal represent better than random classification results, while points below the line represent worse than random results. In general, ROC analysis is one tool to select possibly optimal models and to discard sub-optimal ones independently from the class distribution. Visual examination of ROC curves does not provide quantitative data on the relative classification and predictive performance of multiple algorithms. Area Under Curve (AUC) overcomes this drawback by quantifying the area under the ROC curve, making it easier to find the optimal model.

Results

Chorogram

A chorogram is a graphical representation of the cells of a matrix of correlations. It displays the pattern of correlations in terms of their signs and magnitudes by using visual thinning and correlation-based variable ordering. The cells of the matrix are shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the chorogram in Figure 2, being diagnosed with diabetes has the strongest positive correlation with "HYPEV_A", which indicates whether the respondent has hypertension, and has the strongest negative relationship with "PHSTAT_A", which indicates the respondent's health condition. According to the corresponding responses in Table 1, a lower "HYPEV_A", "DIBEV_A", and "PHSTAT_A" value indicate that the respondent once had hypertension, diabetes, and has a poor general health condition. Thus, people with hypertension and in poor physical condition are more likely to be diagnosed with diabetes.

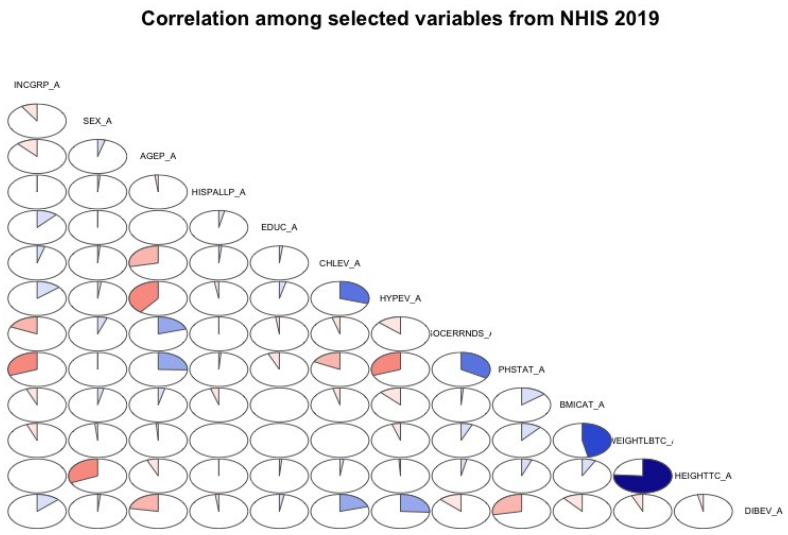

Figure 2: Correlation among variables

Logistic Regression Results

The results of logistic regression analysis of adults ever have diabetes are listed in Figure 3.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.72151    0.17467  21.306  < 2e-16 ***
INCGRP_A      0.24507    0.08094   3.028  0.00246 **
SEX_A         0.61095    0.38544   1.585  0.11295
AGEP_A       -1.76277    0.16324 -10.799  < 2e-16 ***
HISPALLP_A   -0.43127    0.19275  -2.238  0.02525 *
EDUC_A        0.25153    0.37118   0.678  0.49800
CHLEV_A       6.23055    0.48990  12.718  < 2e-16 ***
HYPEV_A       7.06583    0.54412  12.986  < 2e-16 ***
SOCERRNDS_A   0.46731    0.31729   1.473  0.14080
PHSTAT_A     -4.45658    0.24080 -18.508  < 2e-16 ***
BMICAT_A     -1.83504    0.21663  -8.471  < 2e-16 ***
WEIGHTLBTC_A  0.50672    0.20384   2.486  0.01292 *
HEIGHTTC_A   -0.76404    0.23501  -3.251  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Logistic regression results

From the logistic results, it is readily apparent that family income, age, hypertension, cholesterol, health condition, weight and height are significant predictors of having had or likely to be diagnosed with diabetes at the 99% confidence level.

Artificial Neural Network Results

The structure of the artificial neural network is shown in Figure 4. The number on the arrow represents the corresponding weight of that variable.

Garson described a method to identify the relative importance of independent variables for a single dependent variable in an artificial neural network (12). The relative importance of a specific independent variable for the dependent variable can be determined by identifying all weighted connections between the nodes of interest. That is, all weights connecting the specific input node that pass through the hidden layer to the dependent variable are identified. This is repeated for all other independent variables until a list of all weights that are specific to each independent variable is obtained. Figure 5 shows the importance of each question using Garson’s algorithm. The most important predictor of diabetes was hypertension, followed by age, weight, and cholesterol, in decreasing order.

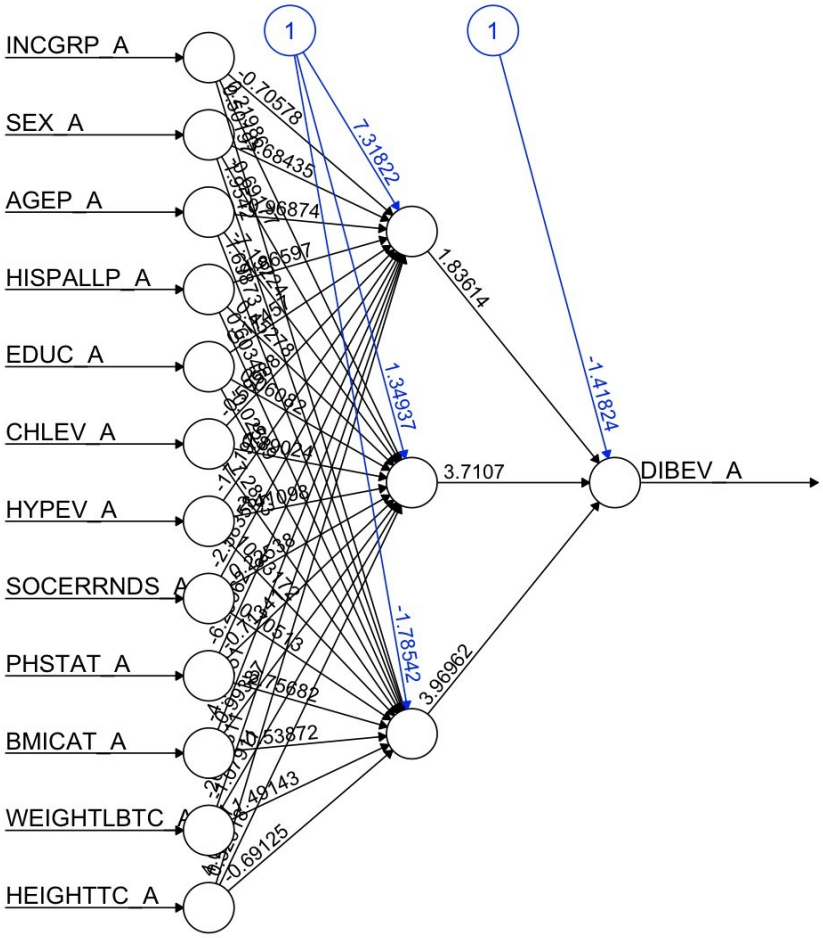

Figure 4: Structure of the artificial neural network

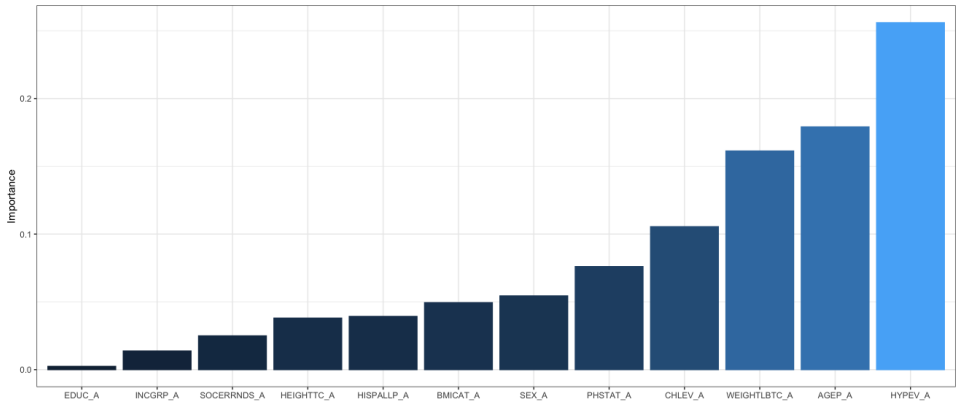

Figure 5: The importance of each question in the artificial neural network.

Model Validation

Figure 6 displays the ROC curve for the Logistic regression and the Artificial neural network models. The AUC scores for the two models were 0.84 and 0.85 respectively. Both models therefore have similar sensitivity and selectivity in predicting diabetes.

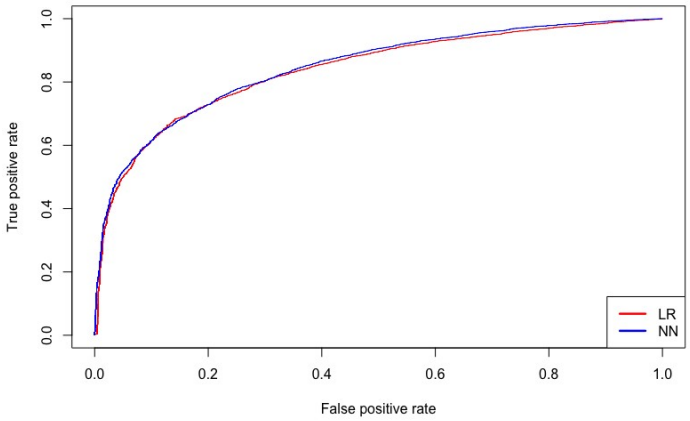

Figure 6: The ROC curve for the logistic regression (LR) and artificial neural network (NN)

Discussion

The purpose of this study was to build a predictive model with the best performance and to identify the factors that would be the most predictive of becoming diabetic in the future. Two models – a logistic regression and an artificial neural network - were constructed. The artificial neural network and the logistic regression models both achieved similar sensitivity and selectivity based on the AUC under the ROC. Using the chorogram and Garson's algorithm, it was concluded that the variable "HYPEV_A" (people who have hypertension) has the highest importance among all independent variables and has the most positive correlation with diabetes, whereas age and health status are negatively correlated. Also, from Figure 3, it can be inferred that income level, age, race, physical state, height, weight, body mass index (BMI), cholesterol, and hypertension are significant predictors of diabetes. In addition, Figure 3 corroborates our finding by showing that hypertension and physical health condition have the strongest positive and the strongest negative correlation respectively of being diagnosed with diabetes. Referring to Table 1, this means that people who have hypertension are more likely to be diagnosed with diabetes and the older and worse the health condition of an individual, the more likely it is that the person has diabetes, or will become diabetic.

We constructed two predictive models in the study. We found that there was no significant

difference between the AUC scores for the artificial neural network and the logistic regression model, for predicting accuracy with an AUC score of 0.85 and 0.84, respectively. This implies that we can apply the same network in future research to help diagnose the probability of diabetes in the public. However, the models have limitations. They are still not sensitive or specific enough; for example, from the ROC curve in Figure 6, the models can correctly diagnose 80% of individuals who are diabetic (sensitivity), with 70% specificity (30% incorrectly diagnosed as being diabetic). As mentioned earlier, this limitation may be due to not resorting to the use of biomarkers. Another limitation is the existence of subjective variables such as physical condition, which may lead to increased bias in the predicted results. Yet another limitation is that using mean value imputation may lead to inaccuracy and bias. To reduce the bias, in a future study, we may use more advanced techniques such k-nearest neighbors (k-NN) algorithm for imputing the missing values. This technique replaces missing values with the mean of *k* (a constant selected by the user) nearest neighbors of that sample, which generally leads to better performance.

Conclusion

A logistic regression model and an artificial neural network were constructed to predict diabetes based solely on responses to the 2019 National health interview survey. Both models could predict diabetes with reasonable sensitivity and selectivity with the AUC scores from their ROC curves being 0.85 and 0.84 respectively. Hypertension and physical health had the strongest positive and the strongest negative correlation respectively with the diagnosis of diabetes. If these models are refined or newer ones constructed so as to reach clinically acceptable sensitivity and selectivity in the absence of biomarker data, that would represent a significant advance toward the diagnosis and treatment of diabetes.

References

1. Boyle, J. P., Honeycutt, A. A., Narayan, K. M. V., Hoerger, T. J., Geiss, L. S., Chen, H., & Thompson, T. J. (2001, November 1). *Projection of DIABETES burden THROUGH 2050*. Diabetes Care. https://care.diabetesjournals.org/conten2t/24/11/1936#:~:text=RESULTS%E2%80%94The%20number%20of%20Americans,%2B437%25%20in%20men).

2. Diabetes - long-term effects. Diabetes - long-term effects - Better Health Channel. (n.d.). https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.

3. Ofei, F. (2005, September). Obesity - a preventable disease. Ghana medical journal. Retrieved February 23, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1790820/

4. Bird, Y., Lemstra, M., Rogers, M., & Moraros, J. (2015, October 12). *The relationship between SOCIOECONOMIC Status/income and prevalence of diabetes and associated conditions: A Cross-sectional population-based study in Saskatchewan, Canada*. International journal for equity in health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4603875/.

5. Hwang, J., & Shon, C. (2014, August 1). *Relationship between socioeconomic status and type 2 diabetes: Results from Korea national health and Nutrition Examination Survey (knhanes) 2010–2012*. BMJ Open. https://bmjopen.bmj.com/content/4/8/e005710.

6. Adler, A. (2021, May 19). Using machine learning techniques to identify key risk factors for diabetes and undiagnosed diabetes. arXiv.org. Retrieved February 23, 2022, from https://arxiv.org/abs/2105.09379

7. Centers for Disease Control and Prevention. (2019, August 5). *NHIS - Questionnaires*. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhis/quest_doc.htm.

8. Cramer. (2002, November). Graduate School and Research Institute in Economics, econometrics and finance. Tinbergen Institute. Retrieved February 24, 2022, from https://www.tinbergen.nl/home

9. Kay, A. (2001, February 12). Artificial Neural Networks. Computerworld. Retrieved February 24, 2022, from https://www.computerworld.com/article/2591759/artificial-neural-networks.html

10. Stefan Fritsch, Frauke Guenther and Marvin N. Wright (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

11. Google. (2020 Aug. 11). *Classification: ROC curve and Auc | machine Learning crash course*. Google. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

12. Garson, G.D. 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6(4):46-51.