Predicting Diabetes with a Logistic Regression Model: A Comprehensive Analysis of NHANES 2011–2012 Data Using SAS

Prince Agyapong

May 2024

1 Introduction

Diabetes occurs when the level of glucose (sugar) in your blood is excessively high. This condition emerges either due to insufficient insulin production by the pancreas or the body's inadequate response to insulin [1]. While diabetes is predominantly a chronic condition, it can be effectively managed through medication and lifestyle modifications [2]. Thus, an intriguing question is the extent to which statistical methods can be used in predicting diabetes. While Meng et al. (2013) [3] evaluate the differences between logistic regression, artificial neural networks, and decision tree models in predicting diabetes using an experimental data set, Dinh et al. (2019) [4] evaluate this effectiveness using the National Health and Nutrition Examination Survey (NHANES) data from 1999-2014.

Our project endeavors to employ logistic regression, using data from the NHANES, to ascertain whether the effectiveness of logistic modeling in predicting diabetes is statistically different when extrapolated to a specific time period (2011-2012). Due to the health shocks immediately following the Great Recession (Margerison-Zilko 2016) [5], these results are a necessary "check" on the preexisting literature. Additionally, we experiment with various bundles of independent variables, providing an expansive set of tools for real-world application. While this

study does not generate new statistical methods for evaluating diabetes, we believe it serves a positive contribution to the preexisting literature.

2 Data Acquisition and Preprocessing

The study utilizes data from NHANES 2011-2012, a comprehensive survey that provides insights into the health and nutritional status of the U.S. population, consisting of 9,756 observations with 8,315 missing[6]. Our analysis incorporates a variety of variables that are potential indicators of diabetes risk:

#	Variable	Type	Len	Label
1	SEQN	Num	8	Respondent sequence number
2	RIAGENDR	Num	8	Gender
3	RIDAGEYR	Num	8	Age in years at screening
4	LBDHDD	Num	8	Direct HDL-Cholesterol (mg/dL)
5	LBDLDL	Num	8	LDL-cholesterol (mg/dL)
6	PAQ635	Num	8	Walk or bicycle
7	ALQ130	Num	8	Avg # alcoholic drinks/day - past 12 mos
8	WHD020	Num	8	Current self-reported weight (pounds)
9	WHQ030	Num	8	How do you consider your weight
10	BPQ020	Num	8	Ever told you had high blood pressure
11	DIQ010	Num	8	Doctor told you have diabetes
12	LBXGLU	Num	8	Fasting Glucose (mg/dL)
13	LBXIN	Num	8	Insulin (uU/mL)

Figure 1: List of Model Variables

- **Demographic Information:** Includes age (RIDAGEYR) and gender (RIAGENDR), as basic yet crucial determinants of diabetes risk. RIAGENDR is categorical, which is

coded as a 1 if the individual is a male and 2 if the individual is a female. RIDAGEYR is continuous, and is truncated at 80 years of age.

- **Lipid Profiles:** Levels of High-Density Lipoprotein (HDL, LBDHDD) and Low-Density Lipoprotein (LDL, LBDLDL), which are key to understanding cardiovascular health and its relationship with diabetes. HDL is a continuous variable in the range of 10-189, and LDL is a continuous variable in the range of 18-357.
- **Physical Activity:** PAQ635 captures data on engagement in physical exercise, a key factor in weight management and reducing diabetes risk. The original coding for this variable assigned a value of 1 to individuals who engage in any physical activity and 2 to those who do not. The variable was recoded for clarity: individuals who engage in physical activity retained a value of 1, while those who do not had their value changed from 2 to 0. Refusal to answer, or unsure answers, are entered in as missing.
- **Alcohol Consumption:** Average alcohol consumption (ALQ130) influences glucose metabolism. This is a continuous variable consisting of the number of alcoholic beverages consumed within the past month, with equivalency being drawn between a 12oz. beer. Refusal to answer, or unsure answers, are entered in as missing.
- **Body Weight:** Data on weight (WHD020), directly linked to diabetes risk. This is a continuous variable in the range of 72-484. Refusal to answer, or unsure answers, are entered in as missing.
- **Perception of Body Weight:** If you have ever been pregnant, what did you consider your weight before you were pregnant? 1 is coded as overweight, 2 is coded as underweight, and 3 is coded as ”just about” the right weight. Refusal to answer, or unsure answers, are entered in as missing.
- **Blood Pressure:** High blood pressure (BPQ020) is linked to a higher risk of diabetes, serving as an indicator of overall vascular health. The original coding for this variable assigns a value of 1 for individuals with hypertension and 2 for those without. To better represent the data, the variable was recoded, changing the value of 2 to 0 to clearly indicate that the individual is not hypertensive. Refusal to answer, or unsure answers, are

entered in as missing.

- **Plasma Fasting Glucose and Insulin:** LBXGLU (Fasting Glucose) and LBXIN (Insulin), crucial for understanding the body's ability to regulate blood sugar and manage insulin, respectively, are paramount for predicting diabetes. LBXGLU is a continuous variable coded from 47-451, and LBXIN is a continuous variable coded from .75-485.1.
- **Diabetes:**DIQ010 indicates whether a respondent has been diagnosed with diabetes. Originally, this categorical variable used a value of 1 for individuals with diabetes and 2 for those without. For clarity, the variable was recoded: 1 still denotes the presence of diabetes, but the value of 2 was changed to 0 to indicate the absence of the condition. Refusal to answer, or unsure answers, are entered in as missing.

The first 10 observations of the dataset is below:

First 10 Observations of Data

Obs	SEQN	RIAGENDR	RIDAGEYR	LBDHDD	LBDLDL	PAQ635	ALQ130	WHD020	WHQ030	BPQ020	DIQ010	LBXGLU	LBXIN
1	62161	1	22	41	110	0	.	150	3	0	0	92	18.65
2	62162	2	3	.	.	.	.	.	.	.	0	.	.
3	62163	1	14	44	.	1	.	.	.	.	0	.	.
4	62164	2	44	28	151	0	.	139	3	0	0	82	3.51
5	62165	2	14	63	84	1	.	.	.	.	0	88	15.35
6	62166	1	9	.	.	.	.	.	.	.	0	.	.
7	62167	1	0	.	.	.	.	.	.	.	.	.	.
8	62168	1	6	51	.	.	.	.	.	.	0	.	.
9	62169	1	21	43	73	0	2	120	3	0	0	107	9.64
10	62170	1	15	61	77	0	.	.	.	.	0	99	9.12

Figure 2

This dataset, with its rich and varied parameters, serves as the backbone for our analysis, offer-

ing a nuanced view of the factors influencing diabetes risk.

3 Methodology

Logistic Regression Modelling

At the heart of our approach will lie logistic regression, which permits the use of continuous or categorical predictors and provides the ability to adjust for multiple predictors [7].

Logistic regression begins with the transformation of the dependent variable into a form suitable for modeling binary outcomes. The key to this transformation is the logit link function, the natural logarithm of the odds ratio of the probability of an event occurring.

The odds of an event is a measure of how likely the event is to occur compared to it not occurring. Mathematically, the odds ratio for an event i is defined as $\frac{\pi_i}{1-\pi_i}$ where p_i represents the probability of the event of interest occurring.

The logit (log of odds) transformation is applied to convert the odds ratio into a continuous scale that can extend from negative to positive infinity. The logit for event i is expressed as:

$$\text{logit}_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{1}$$

This logit is modeled using a linear combination of the independent variables:

$$L(p) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} \tag{2}$$

The coefficients $(\beta_0, \beta_1, ..., \beta_n)$ of the logistic regression model are estimated using Maximum Likelihood Estimation (MLE).

MLE is a method of estimating the parameters of a statistical model, which finds the parameter values that maximize the likelihood of making the observations given the parameters. To perform MLE, the likelihood function $L(\beta)$ for the logistic model must be maximized:

$$L(\beta; p_i, y_i) = \prod_{i=1}^{N} \cdot f(\beta; p_i, y_i) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \tag{3}$$

where y_i are the observed binary outcomes (0 or 1), and X_i are independent Bernoulli random variables with p_i, which is the probability of an event occurring, from the sample.
The log-likelihood function, which is often simpler to maximize, is the natural logarithm of the likelihood function:

$$\ell(\beta; X, y_i) = \prod_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \tag{4}$$

More generally, the model cannot be solved theoretically, only in application. Therefore, the estimation of the unknown parameter, $\hat{\beta}$ is given as:

$$\hat{\beta} = argmax_{\beta} \cdot \ell(\beta; X, y_i) \tag{5}$$

The estimation process involves iteratively adjusting the coefficients, β, based on the training data, to find the set that maximizes this log-likelihood function. To evaluate the model, we will utilize metrics such as the ROC-AUC score[1] to verify the model's effectiveness and reliability in making accurate predictions[8].

Data Division for Model Training and Testing

To ensure that our model would perform robustly on unseen data, we divided our dataset into a training set and a testing set. Specifically, the data was randomly split in an 60:40 ratio using simple random sampling, ensuring that 60% of the data was used for model training while the remaining 40% was reserved for final evaluation. This split was chosen to balance the need for sufficient training data with adequate testing to validate the model's generalizability.

[1]The ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC (Area Under the Curve) represents the degree of separability achieved by the model, with higher values indicating better performance and an ideal value of 1.

Tools and Technologies: Utilizing SAS Programming

SAS is employed throughout the entire process, from initial data manipulation and exploration to the modeling and evaluation stages [9]. SAS programming is particularly effective when attempting to merge multiple data sets together, which was the case when examining the NHANES 2011-2012 data.

In the process of evaluating this relationship, multiple data sets were merged together, using various *"if, then"* statements, as shown in Section 5.1. In order to evaluate the efficacy of the merged data sets, *"proc contents"* was used, and in order to manipulate specific variables, *"proc format"* was used. Graphically, there were two different commands that were utilized: *"proc format"* – as shown in Figure 4 and *"proc freq"* – as shown in Figure 6. To create the testing set, *proc surveyselect"* was used, with a sample rate of .6. Finally, in order to employ logistic regression, *"proc logistic"* was used in conjuction with *"descending outmodel"*, where RIAGENDR, PAQ365, and BPQ020 were used as the CLASS, and DIQ010 was used as the dependent variable. The specific code is shown in Section 5.1, including the actual logistic regression, as well as the model validation.

4 Results

Descriptive Statistics

In this section, we present a comprehensive overview of the key variables included in our study. Descriptive statistics offer insights into the distribution and central tendencies of our dataset. The following table summarizes the minimum, maximum, lower quartile, median, and upper quartile values for each of the quantitative variables. This analysis is critical for understanding the range and distribution of important biomarkers and health behaviors in our sample population.

Descriptive Statistics of Quantitative Variables

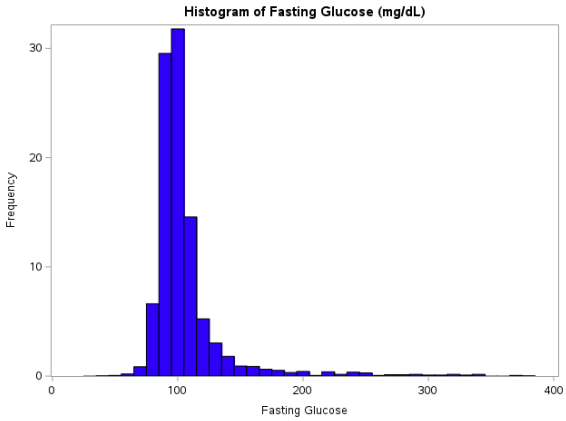
Variable	Label	Minimum	Maximum	Lower Quartile	Median	Upper Quartile
LBDHDD	Direct HDL-Cholesterol (mg/dL)	14.00	175.00	43.00	51.00	60.00
LBDLDL	LDL-cholesterol (mg/dL)	9.00	331.00	84.00	106.00	131.00
ALQ130	Avg # alcoholic drinks/day - past 12 mos	1.00	82.00	1.00	2.00	3.00
WHD020	Current self-reported weight (pounds)	70.00	464.00	140.00	170.00	200.00
LBXGLU	Fasting Glucose (mg/dL)	39.00	382.00	91.00	98.00	107.00
LBXIN	Insulin (uU/mL)	0.14	647.50	6.81	10.63	17.08

Figure 3

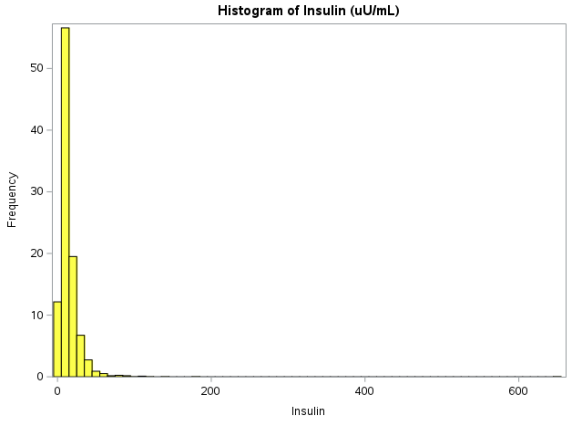
The HDL cholesterol levels range from 14.00 to 175.00 mg/dL, with a median at 51.00 mg/dL, suggesting moderate cardiovascular health among the majority with a fairly narrow interquartile range from 43.00 to 60.00 mg/dL. Conversely, LDL cholesterol shows a broader range from 9.00 to 331.00 mg/dL, with a median of 106.00 mg/dL, reflecting varied lipid profiles and potential for cardiovascular diseases. Alcohol consumption varies drastically, ranging from 1.00 to 82.00 drinks per day, with 75% of participants consuming three or fewer drinks daily, although the high maximum suggests instances of extreme consumption. Participant weights also vary significantly, from 70.00 to 464.00 pounds, with a median at 170.00 pounds, underscoring the presence of both obesity and underweight issues within the study cohort. Fasting glucose levels from 39.00 to 382.00 mg/dL, with a median at 98.00 mg/dL, highlight differing levels of glycemic control, with some values indicating possible diabetic conditions. Lastly, insulin levels range from 0.14 to 647.50 uU/mL, with a median of 10.63 uU/mL, where the upper quartile at 17.08 uU/mL and an extremely high maximum suggest significant cases of insulin resistance or hyperinsulinemia.

We examine the distribution of the variables through histograms as shown in Figure 4.

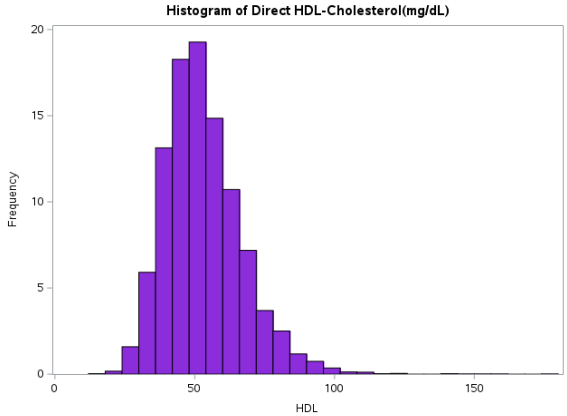
For average alcoholic drinks per day, most individuals consume fewer than 10 drinks, with a steep decline as consumption increases. The fasting glucose histogram indicates a concentration of values around the median, suggestive of consistent metabolic control within the popula-



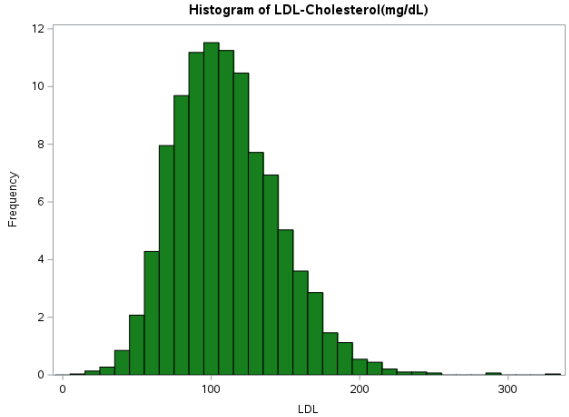
(a) Fasting Glucose levels



(b) Insulin levels

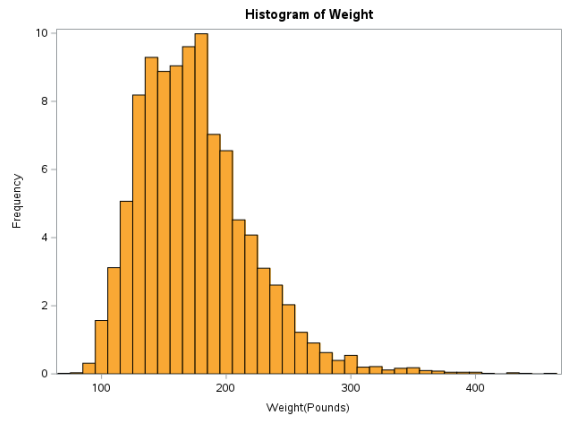


(c) HDL Cholesterol levels

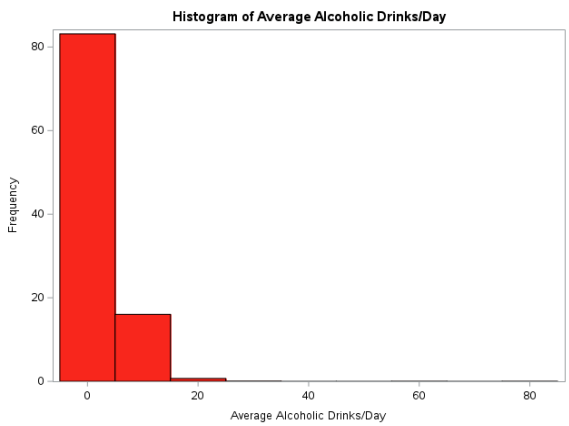


(d) LDL Cholesterol levels

Figure 4: Histograms showing the distribution of metabolic and cholesterol variables.



(e) Weight



(f) Average Alcoholic Drinks/Day

Figure 4: (Continued) Histograms showing the distribution of weight and alcohol consumption variables.

tion, with fewer individuals displaying extremely high or low levels. The distribution of HDL cholesterol is somewhat normally distributed, clustering around the middle range, indicating a balanced lipid profile for the majority of participants. Insulin levels show a high frequency of lower values with a rapid drop-off, reflecting a common baseline in insulin regulation among the participants. LDL cholesterol levels present a relatively symmetrical distribution, peaking around the median value, which may indicate a standardized risk factor for cardiovascular conditions across the cohort. Lastly, the histogram of weight reveals a right-skewed distribution, suggesting that while most participants have a weight within a healthy range, there is a long tail of individuals with higher weights, hinting at potential obesity concerns within the sample.

In examining the relationship between diabetes and key categorical health indicators, a series of tables delineate the incidence of diabetes across different demographic and health behavior groups. These tables highlight how diabetes prevalence varies by gender, levels of physical activity, categorizations of weight, and the status of blood pressure among the study participants.

Percentage of People with Diabetes by Gender

Frequency Percent	Table of DIQ010 by RIAGENDR			
		RIAGENDR(Gender)		
	DIQ010(Doctor told you have diabetes)	Male	Female	Total
	Not Diabetic	4241 45.94	4283 46.39	8524 92.33
	Diabetic	361 3.91	347 3.76	708 7.67
	Total	4602 49.85	4630 50.15	9232 100.00
	Frequency Missing = 524			

Figure 5

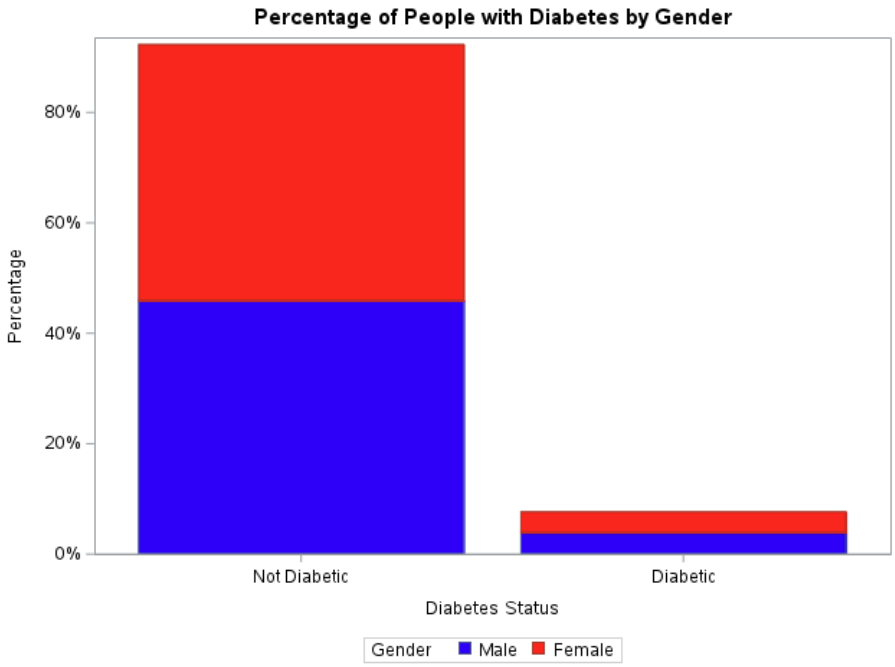


Figure 6: Diabetes by Gender

Percentage of People with Diabetes by Physical Exercise

Frequency Percent	Table of DIQ010 by PAQ635		
	PAQ635(Walk or bicycle)		
DIQ010(Doctor told you have diabetes)	No	Yes	Total
Not Diabetic	3784 56.91	2157 32.44	5941 89.35
Diabetic	559 8.41	149 2.24	708 10.65
Total	4343 65.32	2306 34.68	6649 100.00
Frequency Missing = 3107			

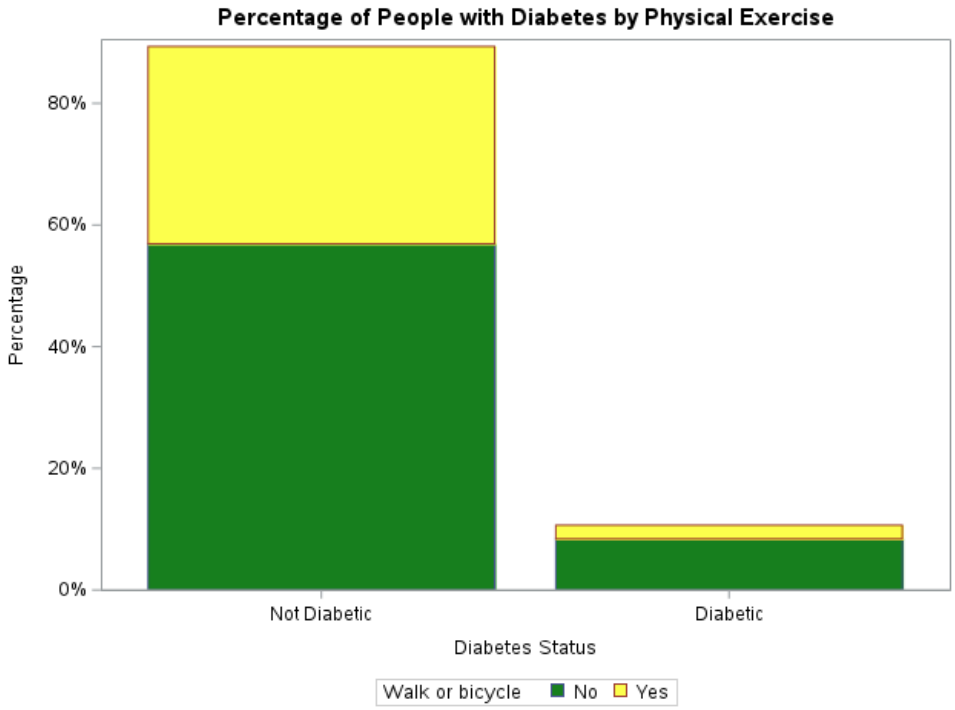


Figure 7: Diabetes by Physical Activity

Percentage of People with Diabetes by Weight Category

Frequency Percent	Table of DIQ010 by WHQ030			
DIQ010(Doctor told you have diabetes)	WHQ030(How do you consider your weight)			
	Overweight	Underweight	Normal	Total
Not Diabetic	2354 39.44	341 5.71	2571 43.08	5266 88.24
Diabetic	437 7.32	25 0.42	240 4.02	702 11.76
Total	2791 46.77	366 6.13	2811 47.10	5968 100.00
Frequency Missing = 3788				

Figure 8

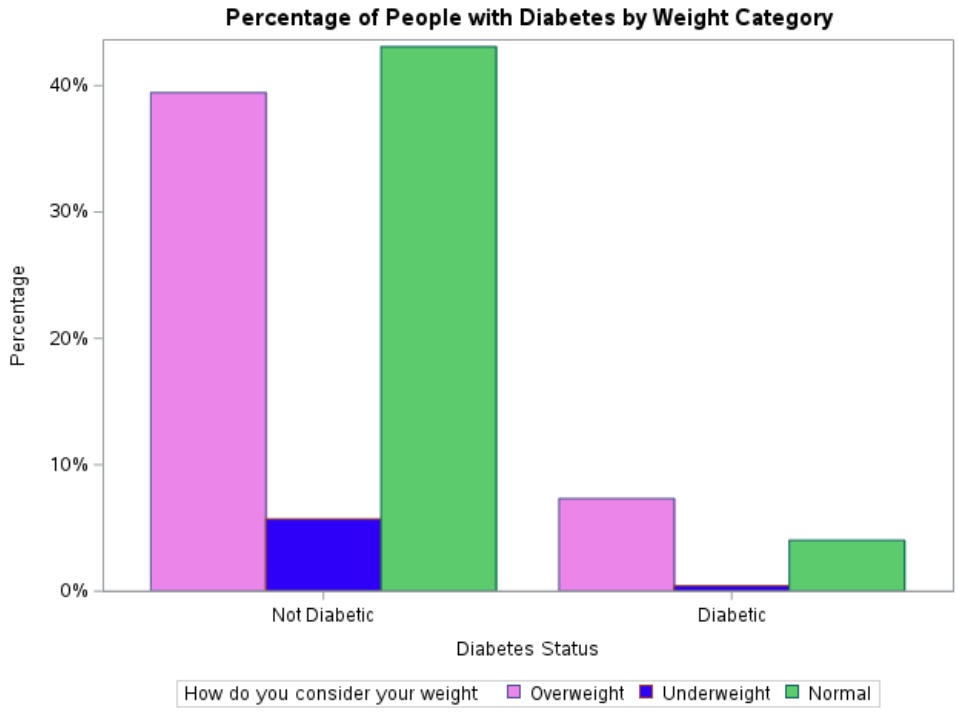


Figure 9: Diabetes by Weight Category

Percentage of People with Diabetes by BP Status

Frequency Percent	Table of DIQ010 by BPQ020		
DIQ010(Doctor told you have diabetes)	BPQ020(Ever told you had high blood pressure)		
	Not Hyptertensive	Hypertensive	Total
Not Diabetic	3869 64.07	1463 24.23	5332 88.29
Diabetic	219 3.63	488 8.08	707 11.71
Total	4088 67.69	1951 32.31	6039 100.00

Frequency Missing = 3717

Figure 10

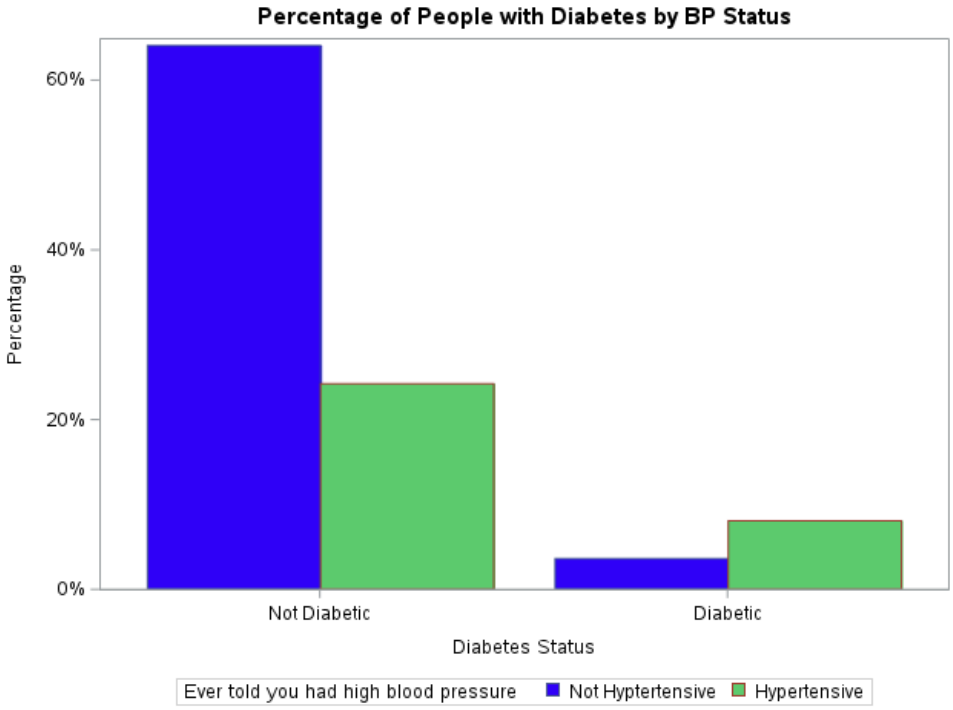


Figure 11: Diabetes by BP Status

Logistic Model

In the analysis of the logistic model using Maximum Likelihood Estimation (MLE), the findings reveal significant relationships between several predictors and the binary outcome. The model's negative intercept (-9.6903) at a highly significant p-value (< 0.0001) suggests that, with all other predictors held at zero, the log odds of the outcome are considerably lower, underscoring the importance of other variables in the model.

Among the predictors, age (RIDAGEYR), LDL cholesterol (LBDLDL), weight (WHD020), and glucose levels (LBXGLU) stand out as significant. Specifically, each additional year of age is associated with an increase in the log odds of the outcome by 0.0479, which is statistically significant $(p < 0.0001)$. This indicates a strong age effect within the model. Conversely, higher levels of LDL cholesterol are linked to a decrease in the likelihood of the outcome, as evidenced by a negative coefficient (-0.0144) with a significant p-value (0.0012). Similarly, each unit increase in waist circumference and glucose levels significantly raises the log odds of the outcome by 0.0138 $(p = 0.0004)$ and 0.0310 $(p < 0.0001)$, respectively, highlighting their predictive relevance.

However, several variables such as gender (RIAGENDR), HDL cholesterol (LBDHDD), insulin levels (LBXIN), physical activity (PAQ635), alcohol intake (ALQ130), and blood pressure concern (BPQ020) did not show significant associations with the outcome in this model.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-9.6903	1.3514	51.4184	<.0001
RIAGENDR	2	1	0.2881	0.1593	3.2697	0.0706
RIDAGEYR		1	0.0479	0.0101	22.3419	<.0001
LBDHDD		1	0.0136	0.00940	2.1039	0.1469
LBDLDL		1	-0.0144	0.00445	10.4315	0.0012
WHD020		1	0.0138	0.00387	12.7088	0.0004
LBXGLU		1	0.0310	0.00396	61.2476	<.0001
LBXIN		1	-0.00817	0.0144	0.3218	0.5705
PAQ635	0	1	0.0267	0.1658	0.0259	0.8721
ALQ130		1	-0.0748	0.0892	0.7041	0.4014
BPQ020	0	1	-0.1742	0.1549	1.2656	0.2606

Assessment of Model Fit in Logistic Regression

Assessing the fit of a model is crucial for ensuring the reliability and accuracy of its predictions. The table below presents the fit statistics for two variations of a logistic regression model: one that includes only the intercept, and another that incorporates both the intercept and additional covariates. These statistics—Akaike Information Criterion (AIC), Schwarz Criterion (SC), also

known as Bayesian Information Criterion (BIC), and Negative Two Log Likelihood ($-2\log L$), and graphically, the ROC curve are fundamental indicators of model performance.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	562.193	382.412
SC	566.956	434.802
-2 Log L	560.193	360.412

Figure 12

- **Akaike Information Criterion (AIC)**: This criterion measures the information lost when a model is used to represent the process generating the data. It strikes a balance between model accuracy and complexity, favoring good fit while penalizing excessive parameterization. The data shows a significant reduction in AIC from 562.193 for the intercept-only model to 382.412 for the model including covariates, indicating a marked improvement in model fit with the inclusion of additional predictors.

- **Schwarz Criterion (SC/BIC)**: Similar to the AIC, the SC assesses model fit with a penalty for the number of parameters, imposing a stricter penalty for model complexity. This is especially useful in scenarios where overfitting is a concern. The SC reduction from 566.956 for the intercept-only model to 434.802 for the model with covariates suggests that the additional parameters are justified, significantly enhancing the model's explanatory power without leading to overfitting.

- **Negative Two Log Likelihood** ($-2\log L$): This statistic quantifies the difference between predicted and observed values, where a lower value signifies a model that more

accurately predicts the observed data. The decrease in $-2\log L$ from 560.193 to 360.412 upon adding covariates demonstrates a substantial increase in the accuracy of the model.

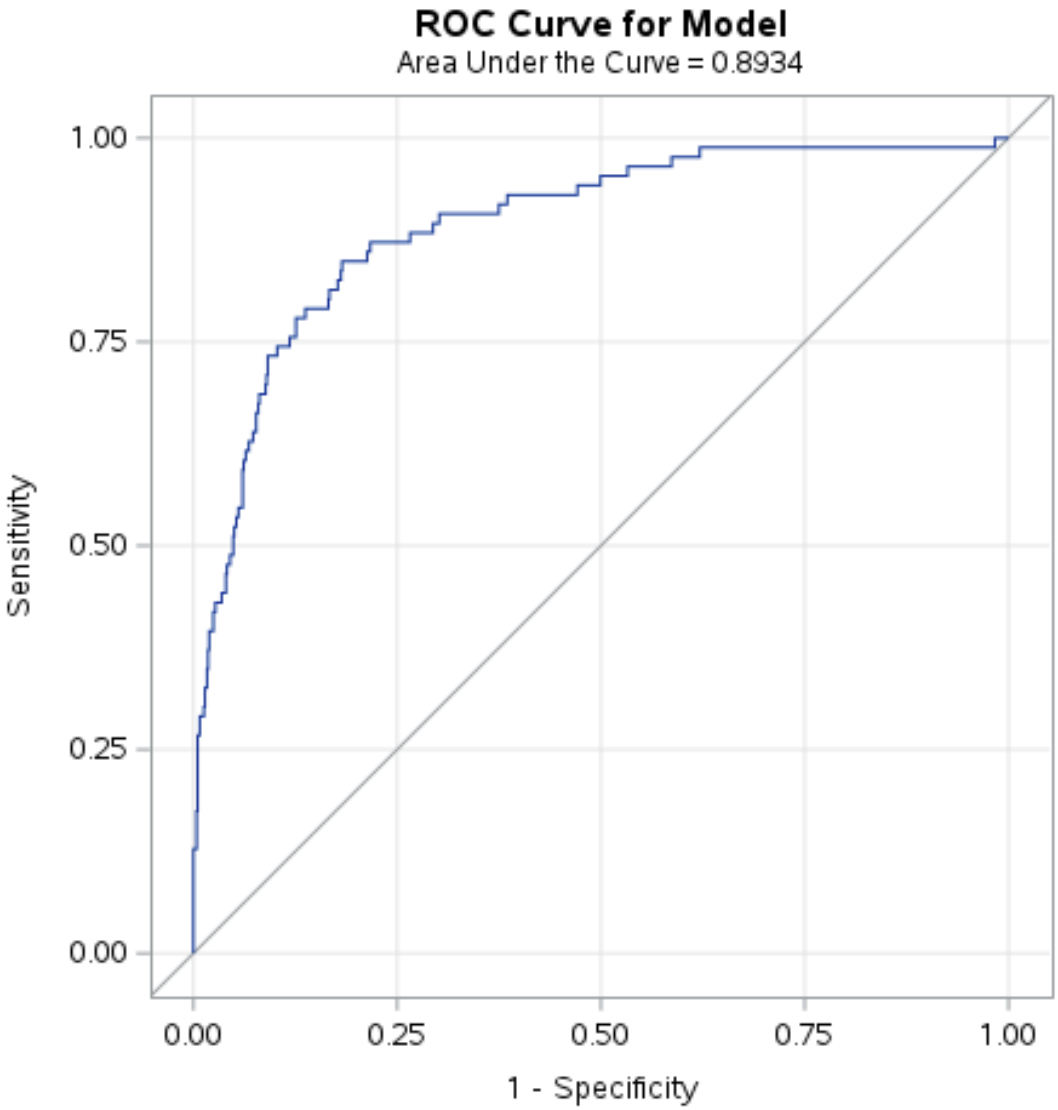


Figure 13: ROC-AUC

The ROC curve displayed in the graph is a performance measurement for the classification model at various threshold settings. The Area Under the Curve (AUC) is 0.8934, which is a strong indicator of the model's ability to distinguish between the two classes.

4.1 Limitations

There are a multitude of limitations to this study. For a large, finite sample, MLE logistic estimation is relatively unbiased. While our study employs a large sample size, the inclusion of

certain variables and the exclusion of certain data points is likely biasing certain coefficients.[2] For example, weight is initially classified as 7777 when the subject refused to answer and 9999 when the subject was unsure of their weight – both of which were reclassified as missing values within our data set. Therefore, the effect of weight is likely more pronounced than that which is present; however, given that this effect is already statistically significant, it is unlikely that the true value of β_{WDH020} is less pronounced or statistically insignificant. Nevertheless, the estimates of β_{ALC130} and β_{ALC130} are likely biased by the missing data, and the significance of their impact on diabetes may be underestimated. Regardless, the AUC score – which is drawn from ROC curve – show that MLE estimates are generally indicative of the true nature of the data.

5 Conclusion

5.1 Implications

In conclusion, our analysis of predictive modeling for diabetes has utilized a robust statistical approach, combining logistic regression with an assessment of model estimates, model fit statistics, and ROC curve analysis to understand and predict the likelihood of diabetes.

We believe that this paper positively contributes to the preexisting literature. While Dinh et al. (2019) [4] provide a more expansive evaluation of this relationship, our paper provides a necessary "check" on the utilization of logistic regression modeling due to the health shocks directly following the Great Recession (Mangerison-Zilko 2016) [5]. The results of this study – and specifically the calculated AUC value under the ROC curve – correspond with similar values found in the literature. While there is certainly more work to be done in the future, this study provides a preliminary examination of successfully utilizing SAS programming – and logistic regression – to model complex health issues. Future research could expand these results over time rather than as a cross-sectional analysis, evaluate these effects in comparison

[2]This problem is typically referred to as attenuation bias. In our study, this could be qualified as self-made attenuation bias. Nevertheless, our reclassification is necessary to employ a logistic model. For the reader, we note that heteroscedasticity is not typically an issue in logistic regression, and is dealt with by the confines of the model.

to the effects of the Covid-19 pandemic (perhaps using difference-in-difference estimation), or examine include new covariates in an attempt to increase the predictive validity of variable-deficient systems.

The logistic regression models indicate that factors such as age, LDL cholesterol, weight, and glucose levels are significant predictors of diabetes, as evidenced by their respective estimates. The model fit statistics with lower AIC, SC, and -2 Log L values for the model with covariates suggest that these predictors contribute meaningfully to the model's accuracy and should be included for a more nuanced prediction.

The ROC curve with an AUC of 0.8934 stands as a testament to the model's discriminative power, signaling a high true positive rate against a low false positive rate. This balance is essential in medical prediction, where the cost of false negatives and false positives carries significant implications.

Appendix A: SAS Source Code


```
proc contents data= projlib.glu_g varnum;
run;

* Demographics;
data work.demo;
set projlib.demo_g;
keep SEQN RIAGENDR RIDAGEYR;
run;

* HDL;
data work.hdl;
set projlib.hdl_g;
keep SEQN LBDHDD;
run;

* LDL;
data work.ldl;
set projlib.trigly_g;
keep SEQN LBDLDL;
run;

* Physical activity;
data work.phyact;
set projlib.paq_g;
keep PAQ635 SEQN;
if PAQ635 = 7 or PAQ635 = 9 then delete;
if PAQ635 = 2  then PAQ635 = 0;
run;


proc freq data=work.phyact;
    tables PAQ635;
    where PAQ635 in (1, 0);
run;


* Alocohol;
data work.alcohol;
set projlib.alq_g;
keep SEQN ALQ130;
if ALQ130 = 777 or ALQ130 = 999 then delete;
run;


*Weight;
data work.wtg;
set projlib.whq_g;
keep SEQN WHD020 WHQ030;
if WHD020 = 7777 or WHD020 = 9999 then delete;
if WHQ030 = 7 or WHQ030 = 9 then delete;
run;



* Blood Pressure;
data work.bp;
set projlib.bpq_g;
keep SEQN BPQ020;
if BPQ020= 2 then BPQ020 = 0;
if BPQ020 = 7 or BPQ020 = 9 then delete;
run;

* Diabetes;
data work.diabetes;
set projlib.diq_g;
keep SEQN DIQ010;
  if DIQ010 =3 or DIQ010 = 7 or DIQ010 = 9  then delete;
  if DIQ010 = 2 then DIQ010 = 0;
run;

proc freq data=work.diabetes;
    tables DIQ010;
    where DIQ010 in (1, 0);
run;



* Insulin;
```

```
data work.glucose;
set projlib.glu_g;
keep SEQN LBXIN LBXGLU;
run;


* Merge all datasets on SEQN;
proc sort data=work.demo;
by SEQN;
run;
proc sort data=work.hdl;
by SEQN;
run;
proc sort data=work.ldl;
by SEQN;
run;
proc sort data=work.phyact;
by SEQN;
run;
proc sort data=work.alcohol;
by SEQN;
run;
proc sort data=work.wtg;
by SEQN;
run;
proc sort data=work.bp;
by SEQN;
run;
proc sort data=work.diabetes;
by SEQN;
run;
proc sort data=work.glucose;
by SEQN;
run;

data work.merged_data;
    merge work.demo
          work.hdl
          work.ldl
          work.phyact
          work.alcohol
          work.wtg
          work.bp
          work.diabetes
          work.glucose;
    by SEQN;
run;

* Print first 10 obs;
proc print data=merged_data(obs=10);
run;

* Print contents;
proc contents data=merged_data varnum;
run;



**************************;



*****************************;
***************Formats;
* Define a format for the gender variable;
proc format;
    value GenderFmt
        1 = 'Male'
        2 = 'Female';
run;

* Define a format for the diabetes variable;
proc format;
    value DiabFmt
        1 = 'Diabetic'
        0 = 'Not Diabetic';
run;
```

```
* Define a format for the ALCOHOL30 variable;
proc format;
    value WHQ030Fmt
        1 = 'Overweight'
        2 = 'Underweight'
        3 = 'Normal';
run;

proc format;
value BPFmt
    1 = 'Hypertensive'
    0 = 'Not Hyptertensive';
run;

proc format;
value PAQFmt
    1 = 'Yes'
    0 = 'No';
run;



* EDA;
* Summary statistics;
proc means data=merged_data min max q1 median q3 maxdec =2;
    var LBDHDD LBDLDL ALQ130 WHD020 LBXGLU LBXIN;
run;


* Diabetes by gender
* Frequency distributions;
title "Percentage of People with Diabetes by Gender";
proc freq data=merged_data;
 format RIAGENDR GenderFmt.;
     format DIQ010  DiabFmt.;
    tables DIQ010*RIAGENDR / norow nocol out=DG;
    ODS NoProctitle;
run;



* Creating a stacked vertical bar chart for Diabetes with formatted gender labels;

proc sgplot data=DG;
    styleattrs datacolors=(blue red);
    vbar DIQ010 / response=count group=RIAGENDR groupdisplay=stack stat=percent;
    xaxis label='Diabetes Status';
    yaxis label='Percentage';
run;


* Diabetes by physical exercise
* Frequency distributions;
title "Percentage of People with Diabetes by Physical Exercise";
proc freq data=merged_data;
 format PAQ635 PAQFmt.;
   format DIQ010  DiabFmt.;
    tables DIQ010*PAQ635 / norow nocol out=DP;
    ODS NoProctitle;
run;


* Creating a stacked vertical bar chart for Diabetes with formatted physical exercise labels;

proc sgplot data=DP;
    styleattrs datacolors=(green yellow);
    vbar DIQ010 / response=count group=PAQ635 groupdisplay=stack stat=percent;
    xaxis label='Diabetes Status';
    yaxis label='Percentage';
run;

* Diabetes by weight status
* Frequency distributions;
title "Percentage of People with Diabetes by Weight Category";
proc freq data=merged_data;
 format WHQ030 WHQ030Fmt.;
   format DIQ010  DiabFmt.;
    tables DIQ010*WHQ030 / norow nocol out=DA;
```

```
    ODS NoProctitle;
run;


* Creating a stacked vertical bar chart for Diabetes with formatted Weight labels ;
proc sgplot data=DA;
    styleattrs datacolors=(violet blue biyg);
    vbar DIQ010 / response=count group=WHQ030 groupdisplay=cluster stat=percent;
    xaxis label='Diabetes Status';
    yaxis label='Percentage';
run;


* Diabetes by blood pressure status
* Frequency distributions;
title "Percentage of People with Diabetes by BP Status";
proc freq data=merged_data;
 format BPQ020 BPFmt.;
   format DIQ010  DiabFmt.;
    tables DIQ010*BPQ020 / norow nocol out=DBP;
    ODS NoProctitle;
run;


* Creating a stacked vertical bar chart for Diabetes with formatted BP labels;
proc sgplot data=DBP;
    styleattrs datacolors=(blue biyg);
    vbar DIQ010 / response=count group=BPQ020 groupdisplay=cluster stat=percent;
    xaxis label='Diabetes Status';
    yaxis label='Percentage';
run;

* Histograms;
proc sgplot data=merged_data;
    histogram LBXIN / binwidth=10  fillattrs=(color=yellow);
    title 'Histogram of Insulin (uU/mL)';
    xaxis label='Insulin';
    yaxis label='Frequency';
run;

proc sgplot data=merged_data;
    histogram  LBXGLU / binwidth=10  fillattrs=(color=blue);
    title 'Histogram of Fasting Glucose (mg/dL)';
    xaxis label='Fasting Glucose';
    yaxis label='Frequency';
run;

proc sgplot data=merged_data;
    histogram LBDHDD / binwidth=10  fillattrs=(color=blueviolet);
    title 'Histogram of Direct HDL-Cholesterol(mg/dL)';
    xaxis label='HDL';
    yaxis label='Frequency';
run;

proc sgplot data=merged_data;
    histogram LBDLDL / binwidth=10  fillattrs=(color=green);
    title 'Histogram of LDL-Cholesterol(mg/dL)';
    xaxis label='LDL';
    yaxis label='Frequency';
run;

proc sgplot data=merged_data;
    histogram WHD020 / binwidth=10  fillattrs=(color=orange);
    title 'Histogram of Weight';
    xaxis label='Weight(Pounds)';
    yaxis label='Frequency';
run;


* Correlation Analysis;
proc corr data=merged_data;
    var RIDAGEYR LBDHDD LBDLDL WHD020 LBXGLU LBXIN  ALQ130 ;
run;


* Missing vlaues;
data merged_data_clean;
    set merged_data;
    if cmiss(of _all_) = 0;
```

```
run;

* Print contents;
proc contents data=merged_data_clean varnum;
run;


* Create a sample data set with 60% of the records;
proc surveyselect data=merged_data_clean out=sample_data
    method=srs /* Simple Random Sample */
    samprate=0.6 /* 60% for training */
    outall /* Outputs all records, selected or not */
    seed=12345; /* Seed for reproducibility */
run;

* Split data into training (60%) and testing (40%) sets;
data training testing;
    set sample_data;
    if Selected = 1 then output training;
    else output testing;
run;

proc contents data=testing;
run;

proc contents data=training;
run;



*Logistic Regression on merged_data_clean with DIQ010 as the outcome;
proc logistic data=training descending outmodel=LogitModel;
    class RIAGENDR (ref='1')
          PAQ635 (ref='1')
          BPQ020 (ref='1');
    model DIQ010(event='1') = RIAGENDR RIDAGEYR LBDHDD LBDLDL WHD020 LBXGLU LBXIN PAQ635 ALQ130 WHD020 BPQ020;
    roc;
run;



* Validate the model on the testing set;
proc logistic inmodel=LogitModel;
    score data=testing out=testing_pred;
run;

* Print first 20 obs;
proc print data=testing_pred(obs=5);
run;
```

References

[1] Cleveland Clinic. *Diabetes*. https://my.clevelandclinic.org/health/diseases/7104-diabetes. Accessed: 2024-04-03. 2023.

[2] S. R. Shrivastava, P. S. Shrivastava, and J. Ramasamy. "Role of self-care in management of diabetes mellitus". In: *Journal of Diabetes & Metabolic Disorders* 12.1 (2013), pp. 1–5.

[3] Xue-Hui Meng et al. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors". In: *The Kaohsiung Journal of Medical Sciences* 29.2 (Feb. 2013), pp. 93–99. DOI: 10.1016/j.kjms.2012.08.016.

[4] An Dinh et al. "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning". In: *BMC Medical Informatics and Decision Making* 19.1 (Nov. 2019). DOI: 10.1186/s12911-019-0918-5.

[5] Claire Margerison-Zilko et al. "Health impacts of the Great Recession: A critical review". In: *Current Epidemiology Reports* 3.1 (Feb. 2016), pp. 81–91. DOI: 10.1007/s40471-016-0068-6.

[6] *National Health and Nutrition Examination Survey 2011-2012*. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011. Accessed: 2024-04-03. Centers for Disease Control and Prevention (CDC), 2011.

[7] D. W. Hosmer. *Applied Logistic Regression*. 2nd ed. Wiley, 2000.

[8] S. G. Grigoryev, Y. V. Lobzin, and N. V. Skripchenko. "The Role and Place of Logistic Regression and ROC Analysis in Solving Medical Diagnostic Task". In: *Journal Infectology* 8.4 (2016), pp. 36–45.

[9] SAS Institute Inc. *SAS Software*. SAS Institute Inc. Cary, NC, USA, 2020.