Pima Indians Diabetes Classification

Case Study 3: DS 501

Richard Valente

Data Science, Worcester Polytechnic Institute, Worcester, NJ, USA rcvalente@wpi.edu

1. Abstract ***(Discovery)***

Obesity related illnesses are on the rise around the world. As of 2019 there are 463 million adults with diabetes and the proportion of the population with diabetes is on the rise (Saeedi 2019). This work leverages machine learning in order to predict if hospital patients have diabetes using a variety of data points collected from medical records and blood samples of the patients. This work could allow for earlier and more accurate detection of diabetes. Early detection would allow for preventative medication and lifestyle changes to be implemented before the disease progresses.

2. Methodology

2.1 Dataset ***(Data Prep)***

The dataset was obtained from kaggle and was originally from the National Institute of Diabetes and Digestive and Kidney diseases (Smith 1998). It contains 768 patients, that are females, of Pima Indian descent and are older than 21. Each patient row contains 8 features used to predict the binary outcome column depicting if the patient had diabetes. Below are descriptions of each feature, these descriptions were obtained from Kaggle (UCI-Machine 2016). Additionally histograms for each feature have been generated to show distributions of the feature set and assess class balance.

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skin fold thickness (mm)

Insulin: 2-Hour serum insulin (muU/ml)

BMI: Body mass index (BMI) = $\frac{weight\ in\ kg}{(height\ in\ m)^2}$

DiabetesPedigreeIndex: Diabetes Pedigree function

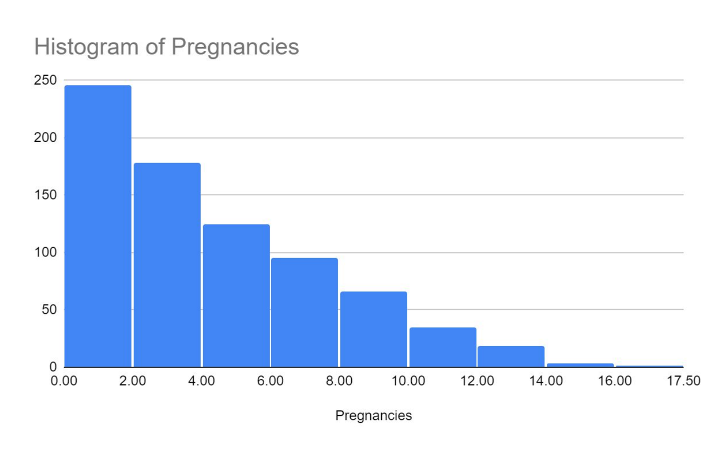
Age: Age (years)

Outcome: Binary outcome of 0 if patient is non-diabetic or one if the patient has diabetes

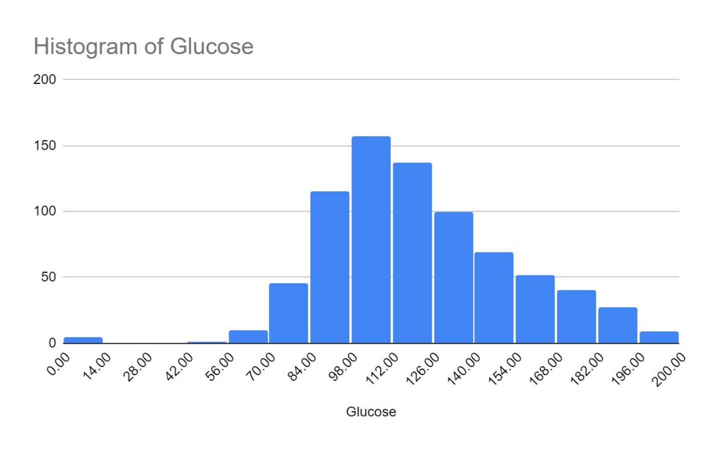
Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0

Table 1. Example of rows in the raw dataset.

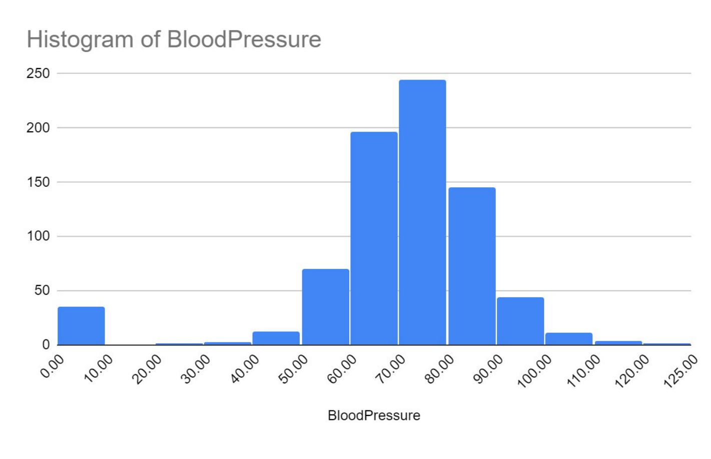
Pregnancies



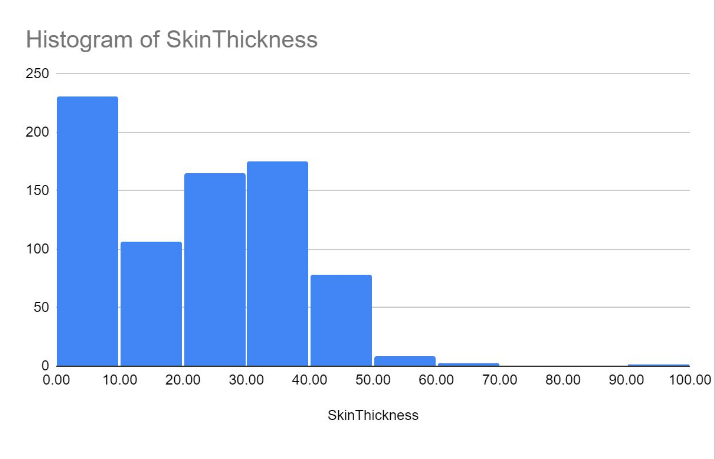
Glucose



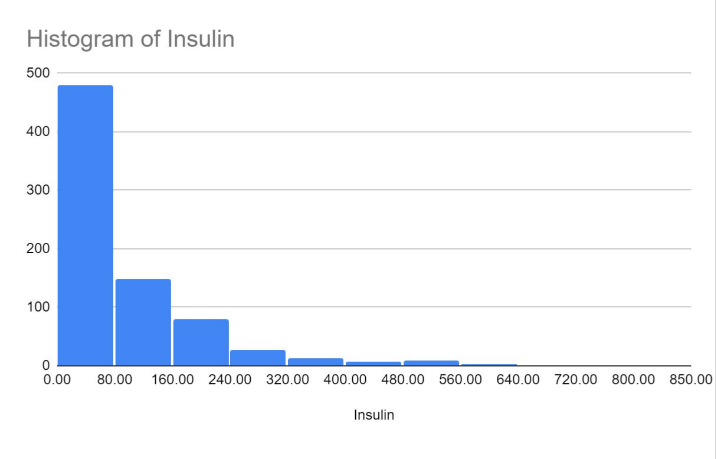
Blood Pressure



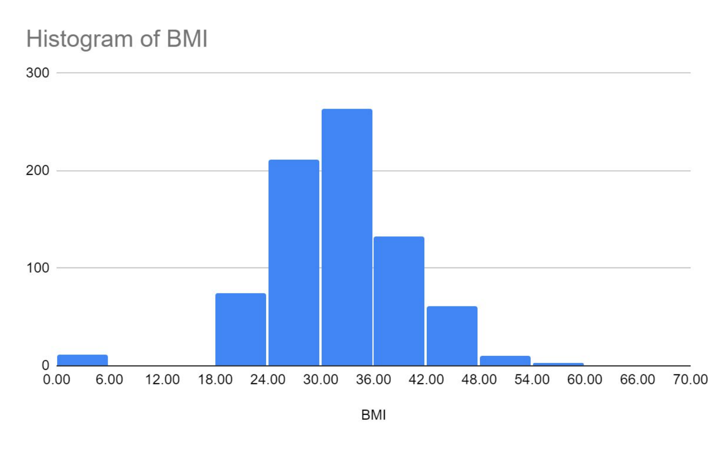
Skin Thickness



Insulin



BMI



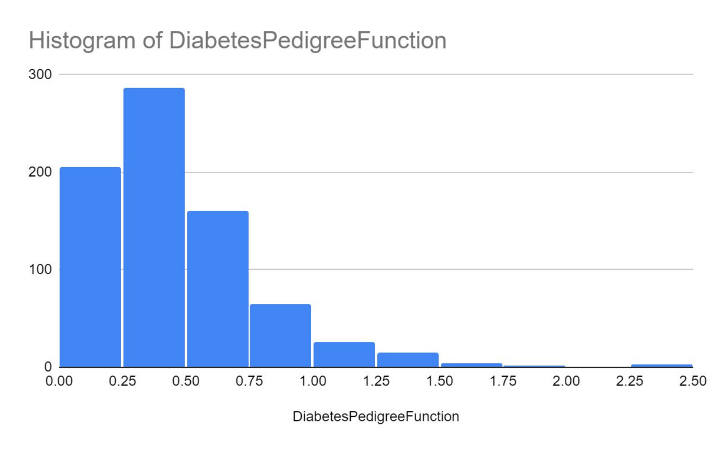
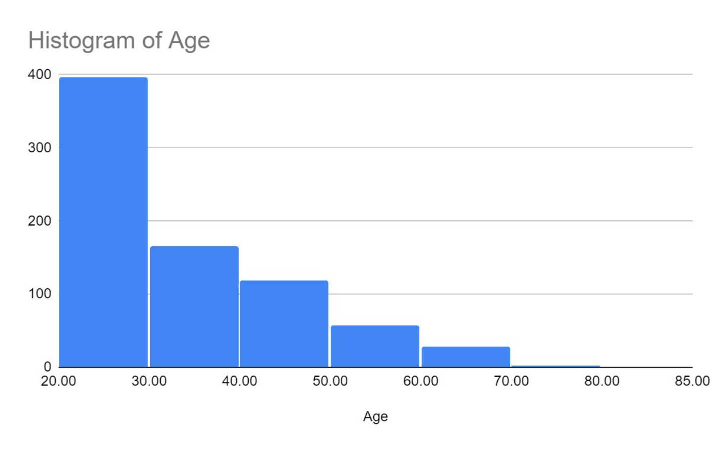
Diabetes Pedigree Function	Age
	

Table 2. Histograms of all predictors in the dataset.

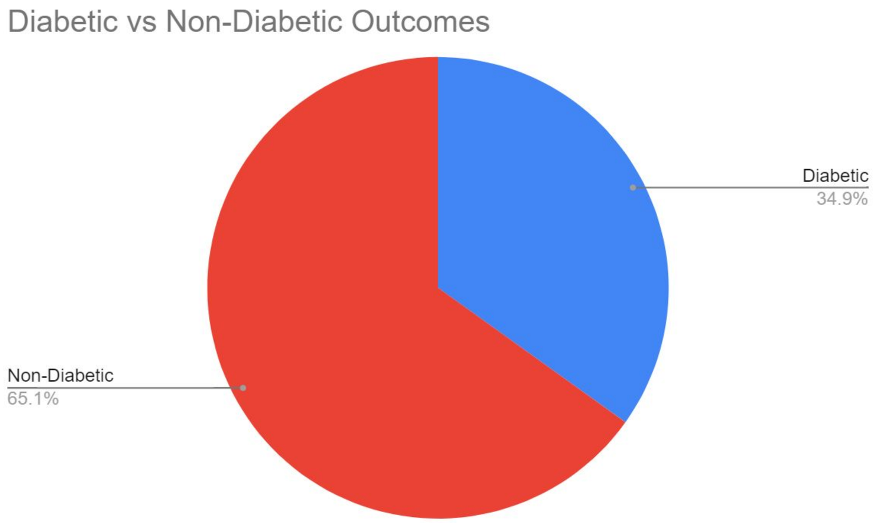


Figure 1. Histogram of outcome response variable if patient had diabetes.

The distributions shown above shows that the featureset is a mix of right-skewed distributions such as ***Insulin*** and ***Pregnancies***. and relatively gaussian distributions such as **Blood Pressure** and ***Glucose*** levels. Data distributions insights show the feasibility of various machine learning classification methods as some have stricter assumptions regarding normalization ect. Additionally the dataset is not completely balanced. There are nearly two times as many Non-Diabetic outcomes compared to Diabetic outcomes. In order to properly assess the accuracy of our model F-score should be used to account for the sensitivity and specificity of the model. Ensuring that the model is not over-fitting.

2.2 Classification Method **(Model Planning)**

In order to predict if patients have diabetes the required predictive model needs to fit the form of the data. This problem will be framed as a binary classification problem where the input predictors will be used to predict the response variable ***Outcome*** which denotes if the patient had diabetes or not. Logistic Regression will be used in order to fit the model to the data.

2.2.1 Logistic Regression

A Logistic Regression model was selected for a few reasons, the dependent variable is discrete. A situation with continuous variables a predictive model such as Linear Regression would be used. Although normalization can improve performance, Logistic regression does not require it as long as extreme outliers are removed from the input data. Logistic regression also assumes a linear relationship between the predictor and dependent variables. If there exists non-linear relationships between these variables a stronger model such as a neural network could improve performance. Logistic Regression is very similar to Linear Regression except with a few important formulas that differentiate it. In order to compare against discrete binary results instead of continuous numerical the Sigmoid Function is the key (Menard 2002).

Sigmoid Function:

$p = \frac{1}{1 + e^{-y}}$

This function converts the output of linear regression into a logit function that maps the values between 0 and 1. This prediction probability can then be used similarly to get a mean-squared-error of the fitted line.

Linear Regression Fit:

$y = b_o + b_1 * x$

Through the transformation of the y variable using the sigmoid function the overall all logistic function is below.

Logistic Function:

$ln(\frac{p}{1-p})$ = $b_o + b_1 * x$

This logistic function can be optimized similarly to linear regression for a minimal R^2 value, through the comparison of the true value y and the predicted value p.

2.5 Training & Validation Process ***(Model Building)***

The dataset was split with a 80% train set and 20% test set. The sets were randomly sampled from the entire dataset. Additionally the RShiny application allows for the model to be built using any set or subset of input parameters to explore effectiveness of various variables. So users can exclude factors such as Skin Thickness.

3. Results ***(Communicate Results)***

The final model had a train accuracy of **78.01%** and validation accuracy of **75.97%**. Below is an overview of key metrics of the binary classification evaluation.

Sensitivity: 86.41%

Specificity: 54.90%

3.1 Summary

Below shows the confusion matrix to showcase the performance of the classifier between positive and negative classifications. As can be seen the performance had a true positive of **79.46%** and a true negative of **66.67%.**

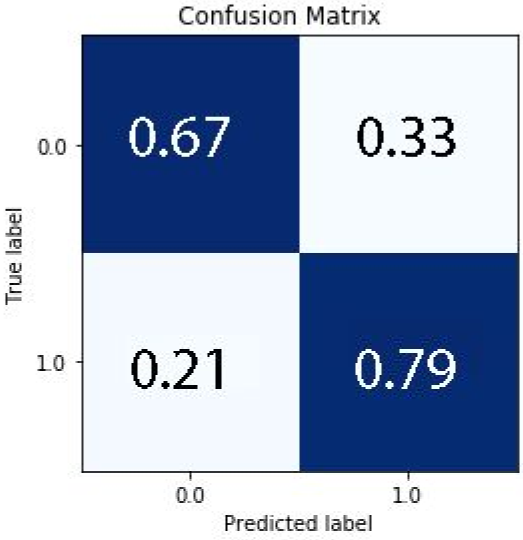


Figure 2. Confusion matrix performanced of final model

These metrics and the confusion matrix ensure two important characteristics of this predictive model. In order to understand the results lets compare them to the results of a naive model that either always predicts the patient has diabetes. The distribution of outcomes is skewed towards the Non-Diabetic patients therefore the performance of such a model would be 65% with a true positive rate of 100% and a true negative rate of 0%. Not only does the trained model outperform this naive model but it also is not “overfitting” as seen by the confusion matrix and the overall sensitivity of the test at detecting diabetes.

3.2 Explainable Predictive Models

Analysis of the significance of the predictors can give insights into the effectiveness of various input parameters. Below is a table showing the p-value of each parameter and its level of significance.

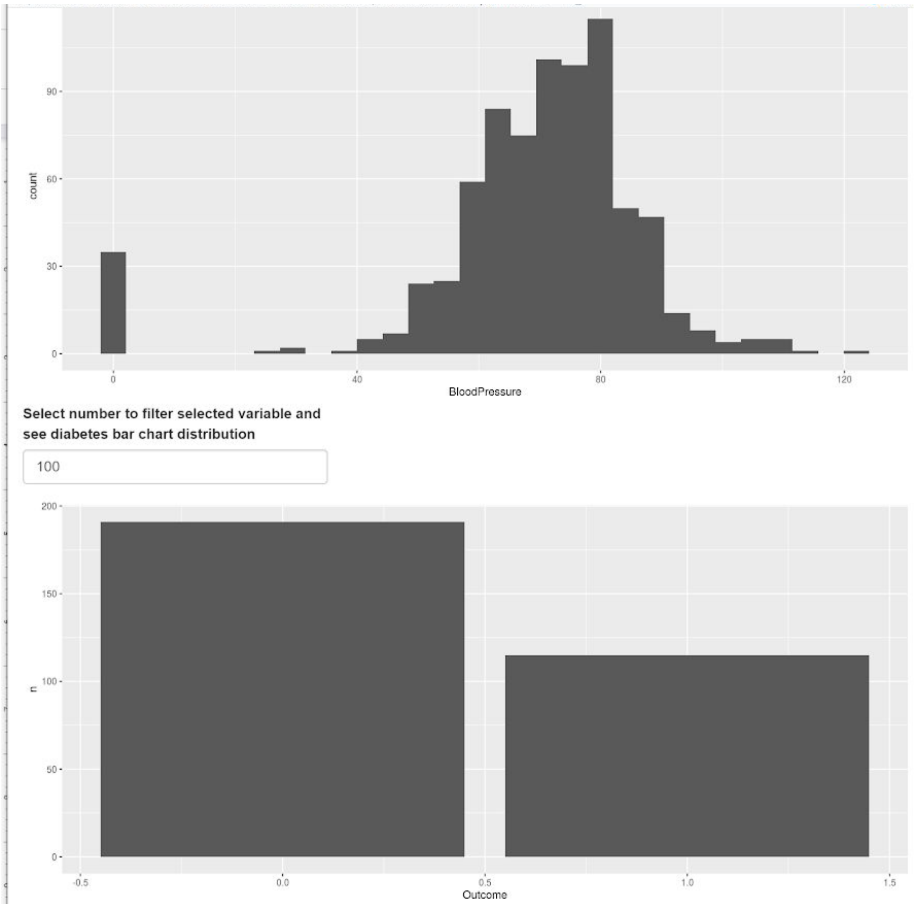
Parameter	P-value	Significance
Pregnancies	< 2e-16	***
Glucose	0.000639	***
Blood Pressure	< 2e-16	***
Skin Thickness	0.0177	*
Insulin	0.684	
BMI	1.26e-07	***
Diabetes Pedigree Function	0.00762	**
Age	0.223	

Table 3. Significance of each predictor in the logistic regression function

The table above shows that the number of Pregnancies and Blood Pressure measurements are the most significant factors in predicting diabetes out of these predictors. Additionally the Insulin levels and Age are not significant factors for the prediction of diabetes. These factors could be removed in future models to improve performance. In order to visualize the significance of these variables the distribution of Diabetic vs Non-Diabetic individuals will be visualized and compared using the created RShiny Application.

4. RShiny Application ***(Operationalize)***

This research has been taken a step further and has implemented a publicly accessible RShiny application accessible at https://rcvalenteai.shinyapps.io/shiny-app/ and on github at https://github.com/rcvalenteai/ds501-hw6 which has a few key features. The application allows users to select a specific parameter to view the histogram plot of that parameters spread, additionally users can filter this variable and see how the distribution of outcomes changes in the represented bar chart.

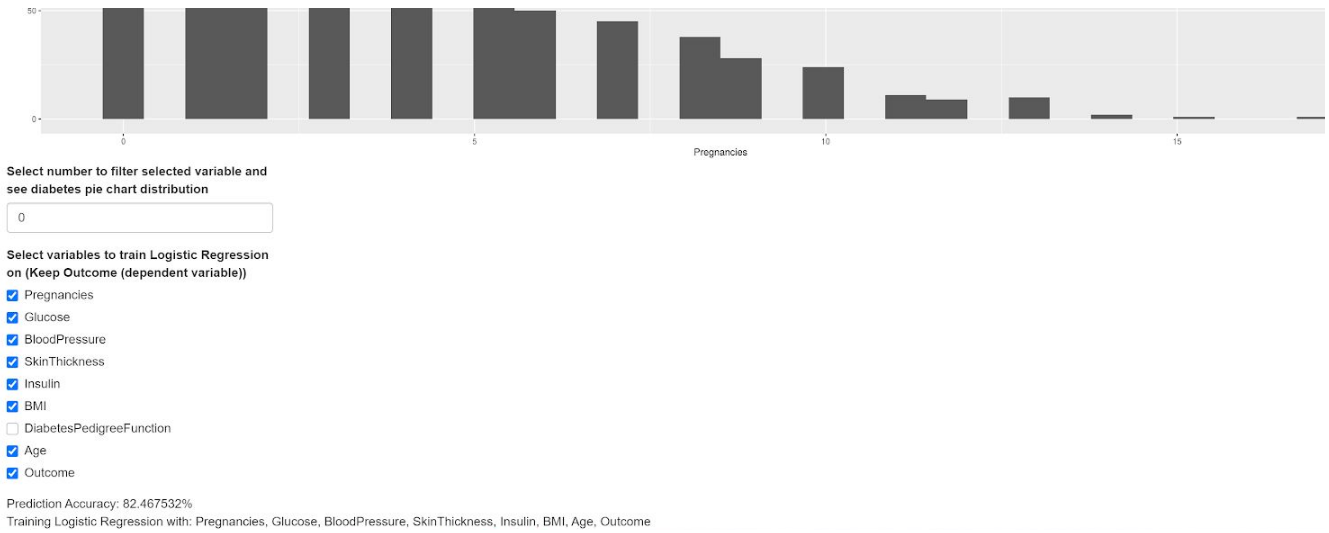


Select number to filter selected variable and see diabetes bar chart distribution

100

Figure 3. Showcase of distribution of outcomes when Blood Pressure is above selected threshold.

In the figure above, the threshold for Blood Pressure was set to show the outcomes of only the patients whose blood pressure was above 90 mmHg. The original distribution changes from a 65% Non-Diabetic to 35% Diabetic balance to a make up of 52% Non-Diabetic to 48% Diabetic balance. This signifies the strong relationship between high blood pressure and diabetes. A number of lifestyle changes and medications can be given to control these symptoms from developing into diabetes. Further exploration can be done through the RShiny application to find more insightful trends.



Select number to filter selected variable and see diabetes pie chart distribution

0

Select variables to train Logistic Regression on (Keep Outcome (dependent variable))

☑ Pregnancies
☑ Glucose
☑ BloodPressure
☑ SkinThickness
☑ Insulin
☑ BMI
□ DiabetesPedigreeFunction
☑ Age
☑ Outcome

Prediction Accuracy: 82.467532%
Training Logistic Regression with: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Age, Outcome

Figure 4. Showcase of parameter input selection for Logistic Regression model training

Additionally researchers can select input parameters for the Logistic Regression training and compare the accuracy to the model trained with all parameters. Since Logistic Regression is negatively impacted by collinearity between variables, removing non significant variables may improve performance. This tool allows researchers to easily tinker, visualize and search the hyperparameters.

5. Conclusion

The overall performance of the model was **75.97%**, this is impressive and shows the value in simple logistic regression. Further work could investigate the use of neural networks to improve accuracy to explore non-linear relationships in the data. Additionally the explainability analysis and RShiny application can be used by stakeholders such as doctors and hospitals analysts to target high risk groups early and apply preventative measures instead of a later dependency on diabetes medicine, specifically insulin which has seen a strain on its supply chain and rapid rise in cost in the past few years (Cefalu 2018).

6. References

Cefalu, William T., et al. "Insulin access and affordability working group: conclusions and recommendations." *Diabetes Care* 41.6 (2018): 1299-1311.

Menard, Scott. *Applied logistic regression analysis*. Vol. 106. Sage, 2002.

Saeedi, Pouya, et al. "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas." *Diabetes research and clinical practice* 157 (2019): 107843.

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

UCI Machine. "Pima Indians Diabetes Database." Kaggle, 6 Oct. 2016, www.kaggle.com/uciml/pima-indians-diabetes-database/data#.