Building Classification Models to Predict Presence of Diabetes

Stephen I. Ellingson

ISYE 6740: Computational Data Analysis

Georgia Institute of Technology

Table of Contents

Project Overview

Diabetes is a chronic disease that originates from abnormally low levels of insulin in the body. Insulin plays an important part in regulating blood sugar levels, and without it, unhealthy amounts of blood sugar can damage nerves or blood vessels over time (Diabetes, n.d.).

According to the World Health Organization, diabetes was in the top ten causes of death with about 1.5 million passing away due to the condition (Diabetes, n.d.). This highlights the ever-important task of having a proactive system to identify when an individual may be at risk of having diabetes. What we are attempting to achieve in this project is using a multitude of classification models learned throughout the ISYE 6740 class, reinforcing the underlying model structure before identifying and comparing accuracy rates. In completing this project, we hope to showcase a holistic understanding of the nuances of different classification models and explain how the power of predictive data may contribute to saving lives.

Background Information

According to information from the CDC, 37.3 million people currently have diabetes - a staggering 11.3% of the United States population (National Diabetes Statistics Report, 2022). About 23% of those adults happen to be undiagnosed, and 96 million people aged 18 or older have prediabetes (National Diabetes Statistics Report, 2022).

These surprising figures are further compounded by the fact that there is an upward trend in prevalence of diabetes. According to a time series analysis conducted by the CDC, the age-adjusted percentage of individuals in the U.S. rose from approximately 7.5% to almost 10% in the span of the last 20 years (Prevalence of diabetes, 2021).

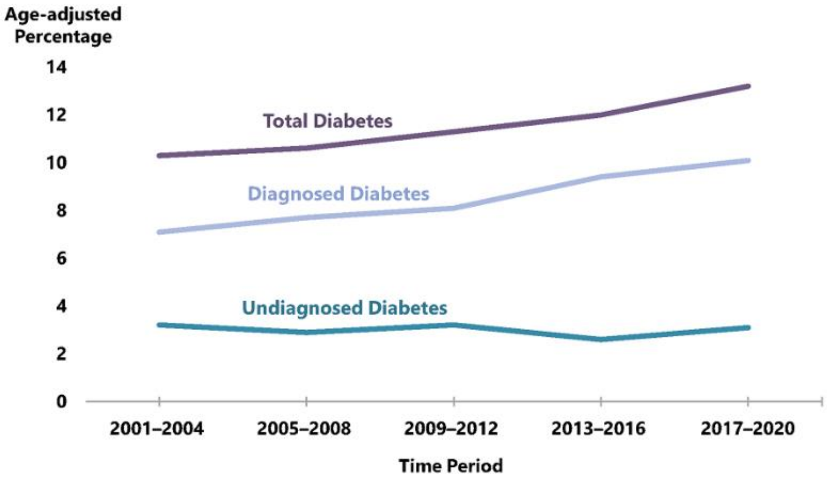


Figure 1: Prevalence of Diabetes in U.S from 2001-2020

Problem Statement

The United States in particular is struggling to manage an ever-growing health concern that is affecting many Americans in present day. Undiagnosed diabetes is another obstacle needing to be overcome, as well. Without proper awareness on the part of individuals afflicted with diabetes, health may worsen over time due to improper treatment. What predictive modelling can offer to mitigate the issues presented by diabetes is assist doctors in diagnosing patients, helping patients understand whether they may be at risk or not, and even aid government agencies in allocating funding and resources to the right areas.

Initial Hypothesis

Before we begin testing multiple classification models, we would like to make a few initial hypotheses. Our first is that logistic regression will provide the most actionable results. As logistic regression cleanly predicts probabilities in addition to making classifications against a certain threshold, we feel that while accuracy may or may not be higher than our other models, the intuitive results will prove that logistic regression will be the best model to use.

Our second hypothesis is related to the data itself. As further discussed below, we have a small dataset of 768 rows with eight predictor variables and one binary response variable. In viewing all predictor values, we hypothesize that insulin levels will be the most significant variable in evaluating presence of diabetes.

Data Overview

We leveraged a dataset originally from the National Institute of Diabetes and Digestive and Kidney Diseases (found in Kaggle) that consists of eight personal key indicators of diabetes. The Kaggle author has segmented this dataset from the metadata to include only females of Indian heritage that are at least 21 years of age.

Data Source

As stated above, the dataset itself was obtained through Kaggle and was originally collected by an American diabetes research organization. It contains 768 data points, with each data point representing a female surveyed.

The feature columns are indicators that can influence the probability of someone having diabetes. We want to analyze this dataset and apply different machine learning methods for classification, prediction, and estimating the likelihood that someone may be at risk of having heart disease.

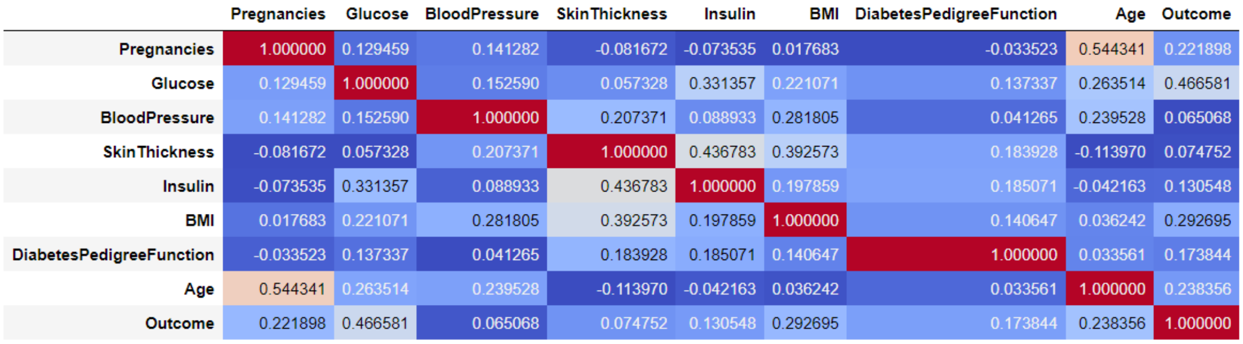
The eight predictor values are described below:

1. Age: Time lived in years of an individual.
2. Pregnancies: Number of times an individual has become pregnant.
3. Diastolic Blood Pressure: Blood pressure levels (mm Hg).
4. Skin Thickness: Triceps skin fold thickness (mm) of an individual.
5. Insulin: Two-hour serum insulin levels (mm U/ml).
6. BMI: Body mass index calculation (kilograms/height (meters)^2).
7. Diabetes Pedigree Function
8. Glucose: Plasma glucose concentration calculated from an oral test.

Data Processing

Since this dataset was obtained from Kaggle, most of the core data processing had already been completed. In addition, since all variables are continuous (excluding the response variable), there is no need to create dummy variables or convert data types. However, there were still several important steps we needed to conduct to transform the data before moving on to modeling:

1. **Scaling/standardization:** When unscaled, relationships between predictors and responses can fail to be accurately captured. Some variable ranges are smaller than others, and because of this, changes in those small-range variables will have a much more drastic effect on response than is truly present. To mitigate this type of error, we scaled all variables to have a mean of 0 and identical standard deviation.
2. **Null values:** These were searched for across the different variables, but none were found.
3. **Duplicate values:** We found no duplicates after initially analyzing the dataset.
4. **Check for correlation:** A final step before modeling involved checking all variables for correlation. After analysis, we determined that no correlation level exceeds a threshold of 0.7. As such, no variables required removal.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Figure 2: Correlation Matrix of All Variables

Methodology

To approach this project, three different phases were defined:

1. **Data preparation:**
   Scale continuous variables, verify no null or duplicate values, check for correlation between predictor variables.
2. **Run Multiple Classification Models:**
   a. **Feature Selection**: Run Lasso regression to identify significant feature variables.
   b. **Model Building**: Create and fit four types of models, with additional model experimentation on logistic regression. The four models and specifications are:
      i. **Logistic Regression:** Four variants, with mixing and matching of:
         1. **Training Methods -** K-fold cross validation, 80/20 training/testing
         2. **Features –** All variables, Lasso regression variables
      ii. **K-Nearest Neighbors**: K-fold cross validation with all variables
      iii. **Linear SVM**: K-fold cross validation with all variables
      iv. **Random Forest**: K-fold cross validation with all variables

   In-depth explanations of both the modelling process and underlying model components are detailed in the following sections.

3. **Model Accuracy Comparison:**
   After building and fitting our models, we compile accuracy rates and compare, with an emphasis on the insightfulness of the data gathered. Afterwards, we identify the most optimal model and explain how results may be used for future research and betterment of the health industry.

Model Overview

As explained in the previous section, we ran a variety of several classification models, as well as a Lasso regression analysis to help us with feature selection. For each model, we will discuss the underlying mathematical structure as well as detailing how necessary parameters were set.

Lasso Regression

The first model employed was not one related to the task of classification. Rather, we wanted to identify the features that most impacted the chance of an individual to have diabetes. As such, we employed Lasso Regression to complete this objective.

With the basic linear regression problem, we are tasked with minimizing the sum of squared error for all data points when fitting a regression line. What Lasso regression adds is a regularization term that controls model complexity, with the regularization term λ varying based on the objective.

$$\theta = argmin_\theta \, L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(y^i - \theta^\top x^i\right)^2 + \lambda||\theta||_1$$

As such, we have a multi-faceted optimization problem in which we want to minimize error and (depending on the regularization term value) maintain model simplicity. The regularization term forces some variable coefficients down towards zero to reduce total error, and thus, we can highlight our significant features.

Using a simple iterative process paired with sklearn's Lasso model function, we tested multiple λ values, each time fitting a regression model to a training dataset, testing the model's performance by viewing R-squared values, and compared results. The model with the highest R-squared value was chosen, and coefficients were identified. We determined a λ value of 0.01 to be the most optimal.

Logistic Regression

While the project's overall focus was to analyze and compare various classification models, a focus was put on logistic regression for its intuitive, simplistic output. In logistic regression, we analyze each data point, sample the value of X_i according to a prior p(x) and sample a label Y_i from the conditional probability function

$$p\left(y \middle| x^i, \theta\right), y = 0, 1$$

This probability function is known as the sigmoid function in the case of logistic regression. Using this information, for any given value of X, we can calculate the probability of X being a certain class. This result is then compared against a predetermined threshold (usually 0.5) to classify.

In our logistic regression analysis, we created four different models, which were combinations of two things: excluding variables removed from Lasso regression and using either cross-validation or an 80/20 training/testing split. The classification threshold was kept at 0.5 for all models for the purpose of simplicity, l2-norm penalty term was used, and no class weights were employed. Results from all four models are discussed in the following section.

K-Nearest Neighbors

With k-nearest neighbors, points are mapped in multiple dimensions, and distances between those points are calculated. The algorithm simply identifies the distance between a new data point and the surrounding k number of data points to predict what class that new data point would fall into. This is decided on a simple majority outcome of the previous step. This implementation uses no neighbor weights (all points have an equal effect on decision) and Euclidean (l2) distance for measurement.

The KNN method can be applied to a labeled classification problem in the following way: train the model on a subset of data and test the predictive value of that model with the remaining data points. Cross-validation may also be used. In this specific implementation, we conducted both k-fold cross validation with a k-value of 10 and an 80/20 data split method, with k-values ranging from 1 to 100. We then plotted the results, as shown in Figure 3. We identified that in this particular instance, a k-value of 3 using the 80/20 split method produced the highest predictive accuracy.

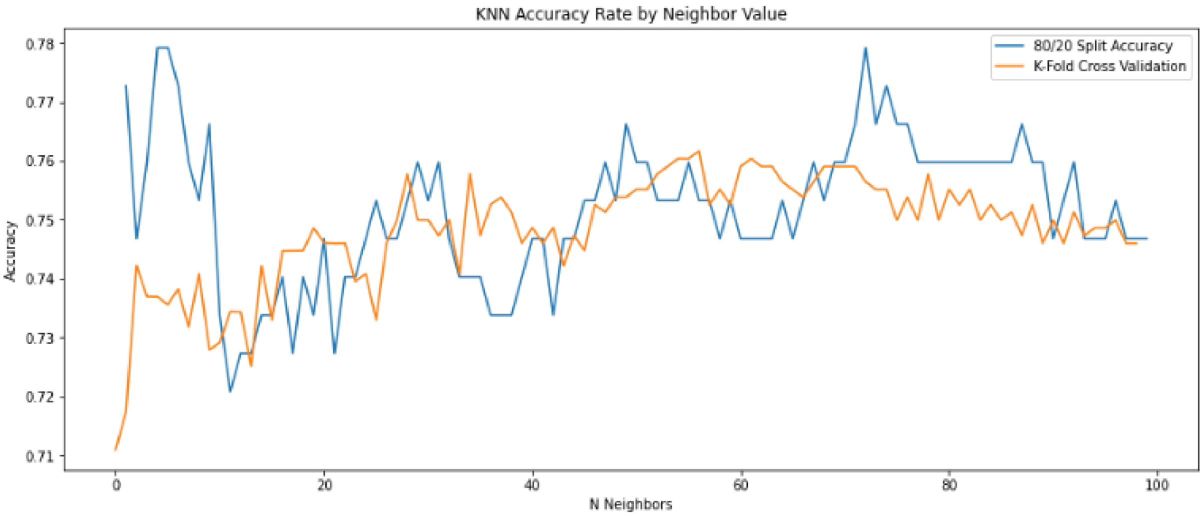


Figure 3: KNN Accuracy Rates

It should be noted with our KNN implementation that while working with labeled data may be considered unorthodox, the algorithm is able to be applied in this fashion. In addition, with this particular dataset, many of the pitfalls of using KNN are avoided, such as lack of robustness against high-dimensional data, large number of data points, and categorical features.

Linear SVM

Like k-nearest neighbor classifiers, support vector machine models use geometric intuitions to make decisions. However, rather than simply analyze point-to-point distances, SVM determines a classification hyperplane between two classes of data points. This hyperplane is calculated by minimizing the training error while also maximizing the distance of the decision plane the two clusters (this is also known as the margin). Depending on the classification problem, we may include error weights for a certain class, meaning that the optimal decision boundary will not lie in an equidistant manner between classes.

In our implementation, we created a linear SVM model in which multiple parameters were tested:

1. Regularization parameter: as stated above, we can specify how much we would like to avoid misclassification. We tested values 0.1, 1, 10, and 100, with higher values corresponding to a decision boundary that is more closely fitted to the data points.
2. Gamma: we may also identify how training points influence the calculation of the decision boundary based on their location from that boundary. We tested values 0.0001, 0.001, 0.01, 0.1, and 1, with low values corresponding to more long-distance points being considered.

With testing our SVM model with each combination of parameters, we were able to identify the optimal result with a regularization value of 1 and a gamma value of 1.

Random Forest

The final model used to predict diabetes data was random forest. As a short review, random forests employ the use of multiple decision trees, aggregating classification outputs for each tree and determining outputs based on average results. Bootstrapping is the sampling method used when calculating this model.

In this iteration, we fit a random forest model using 100 decision trees. Each decision tree was allowed to have no depth limit, so long as each leaf contained at least one sample. Samples were equally weighted when used.

Model Results

Lasso Regression

After conducting initial analysis on our scaled data with Lasso regression, we found that out of the eight feature variables included in the dataset, skin thickness levels were shown to be insignificant in predicting presence of diabetes. Coefficient values can be viewed below.

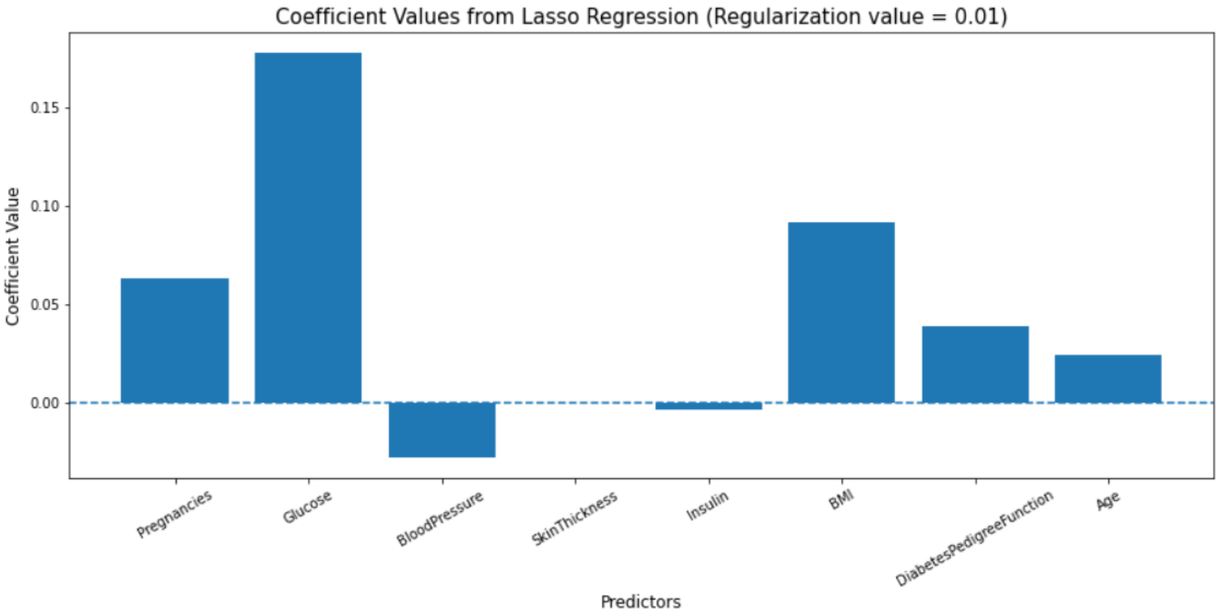
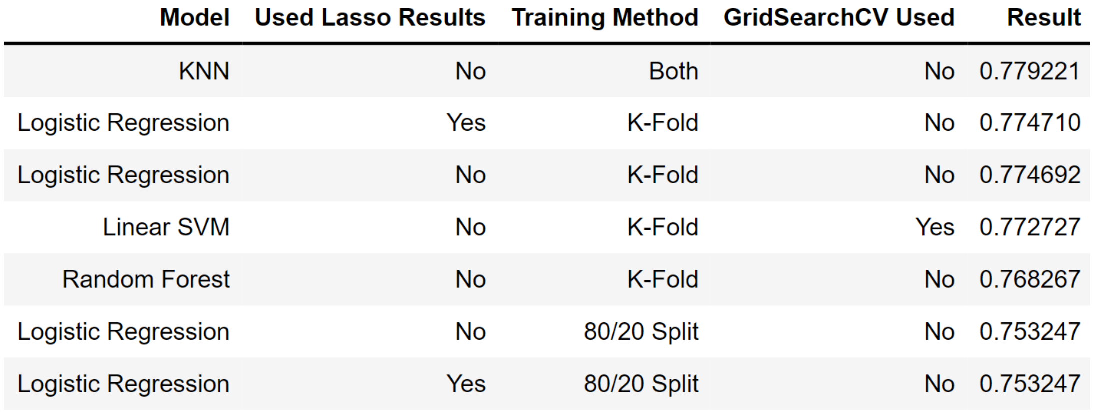


Figure 4: Coefficient Values from Lasso Regression

As depicted above, five of the eight feature variables carry a positive relationship with chance of having diabetes, while two (Blood Pressure and Insulin) have a negative relationship. This intuitively makes sense for many of these variables; for example, obesity has historically been linked with diabetes, which reflects the positive coefficient on BMI.

Classification Models

Using the underlying knowledge of all four of our classification models, as well as the methodology and parameters defined in the previous section, we identified each model's optimal predictive accuracy, as shown below in Figure 5.



Model	Used Lasso Results	Training Method	GridSearchCV Used	Result
KNN	No	Both	No	0.779221
Logistic Regression	Yes	K-Fold	No	0.774710
Logistic Regression	No	K-Fold	No	0.774692
Linear SVM	No	K-Fold	Yes	0.772727
Random Forest	No	K-Fold	No	0.768267
Logistic Regression	No	80/20 Split	No	0.753247
Logistic Regression	Yes	80/20 Split	No	0.753247

Figure 5: Classification Model Results

It was discovered that while all models achieved accuracy rates above 75%, the k-nearest neighbors classifier proved to correctly determine whether or not an individual had diabetes. For the most part, k-fold cross validation was the method of model training that resulted in the highest accuracy rates, as can be seen by the order in the figure above. However, an interesting insight is that our leading model (KNN) achieved its best results using an 80/20 training/testing split.

Model accuracy rates overall were acceptable, and all models ran efficiently due to the dataset size, feature variable types, and preprocessing methods used. When combined with the information gathered from our Lasso regression, we found that including all predictors generally resulted in more accurate models.

Conclusion

The focus of this project was to reproduce concepts learned in a novel approach by holistically executing analysis on real-world data. We ensured proper data collection, cleaning, and scaling, and we identified which model performed best. In the following sections, we detail some key insights gained from this research, discuss the real-world applications of the leading models, and plan next steps for future analysis.

Key Insights

- The K-nearest neighbors model outperformed all other models, including all four logistic regression variants. This disproves our first hypothesis and offers an interesting insight into the combinations of model structure and training to achieve optimal results.
- Insulin levels were largely insignificant in predicting presence of diabetes, according to Lasso regression. This disproves our second hypothesis. Glucose levels were determined to be the feature containing the most predictive information regarding prediction ability.
- Models generally performed better when using all predictor variables, instead of the ones only deemed significant by Lasso regression. This may be due to the fact that there were not many predictor variables to start with, and retaining as much information where applicable is more valuable in this case than minimizing dimensionality.

Identifying the Truly "Best" Model

While no model outperformed the others significantly, we found that KNN topped the list in terms of accuracy. In a real-world application as discussed in the first sections of this project report, KNN would most likely fail to provide as much benefit as a logistic regression model due to its inability to output probabilities. If we were to identify a model that gave more intuitive information to patients, doctors, or policy makers alike, then logistic regression may be the optimal choice for its easily interpretable probabilistic output.

Most Important Predictors of Diabetes

Using Lasso regression, we estimated the most statistically significant indicators to predict diabetes. These were Pregnancies, Glucose, Blood Pressure, BMI, Age, Diabetes Pedigree Function, and Insulin.

Project Challenges

While we didn't encounter major obstacles or challenges to analyze our dataset and determine the best classification model, we did find some observations that, if we were to continue this analysis we would have to consider as next steps. For example, as we looked closer to the impact of correlation between data predictor variables, we noticed that no variables surpassed a correlation threshold of 0.7. However, we found that “Age” and “Pregnancies” were the two predictor variables that shared the closest relationship.

Next Steps

We realize that this project can expand into a useful application for industry professionals, patients, and institutions that would like to better understand the probabilities of being at risk of having diabetes. The following steps for this project could consist of:

- Identifying larger datasets and new predictor variables to analyze with our models.
- Exploring threshold values from the varying logistic regression models created.
- Providing a real-world application in the form of a mobile app or online service to medical officials and the public to easily assess risk of diabetes.

Sources and Dataset

Akturk, M. (2020, August 5). *Diabetes dataset*. Kaggle. Retrieved August 1, 2022, from https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Centers for Disease Control and Prevention. (2021, December 29). *Prevalence of both diagnosed and undiagnosed diabetes*. Centers for Disease Control and Prevention. Retrieved August 1, 2022, from https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html

Centers for Disease Control and Prevention. (2022, January 18). *National Diabetes Statistics Report*. Centers for Disease Control and Prevention. Retrieved August 1, 2022, from https://www.cdc.gov/diabetes/data/statistics-report/index.html

Linear Models. scikit. (n.d.). Retrieved August 1, 2022, from https://scikit-learn.org/dev/modules/linear_model.html#logistic-regression

World Health Organization. (n.d.). *Diabetes*. World Health Organization. Retrieved August 1, 2022, from https://www.who.int/news-room/fact-sheets/detail/diabetes

Xie, Y. (n.d.). Decision Tree and Random Forest.

Xie, Y. (n.d.). Feature (Variable) Selection.

Xie, Y. (n.d.). Introduction to Classification.

Xie, Y. (n.d.). Support Vector Machines (SVM).