Group: Harutyun Hakobyan, Kavi Chikkappa, Alexander G. Sonyey, Hemant Sathian

Professor LuValle

Computation and Graphical Analysis in Statistics 01:960:486 Section 01

08 May 2022

Predicting Diabetes Among Patients

Abstract

Diabetes is a prevalent disease affecting over 450 million people (9.3% of the population) throughout the globe, and it is steadily on the rise with its predicted affected population to be 10.9% by 2045 (Saeedi et al., 2019). With this increased spread of diabetes, predicting the possibility of having the disease becomes all the more important, especially when there are methods of prevention in certain types of diabetes. In fact, one in two (50.1%) people living with diabetes do not know they have diabetes (Saeedi et al., 2019). With a dataset sourced from an Iraqi Hospital Laboratory, we utilized a variety of statistical methods and R functions to develop a strong predictor model which we would later use to confidently predict patients who have the possibility of having diabetes as well as predict potential warning signs of someone being pre-diabetic.

Introduction

Subject Background

When someone is diagnosed with diabetes, insulin is not used as efficiently as it should and/or the body is not able to make enough to sustain itself. Over time, diabetes has a significant effect on a person's ability to use food for energy efficiently. Food that would have been mostly used as an energy source remains in the bloodstream as glucose. Apart from energy generation, diabetes can cause many other complications in a person's body, putting their lives in more danger. Some complications include heart and blood vessel disease, nerve damage in limbs, kidney diseases, eye damage, skin conditions, sleep apnea, and many others.

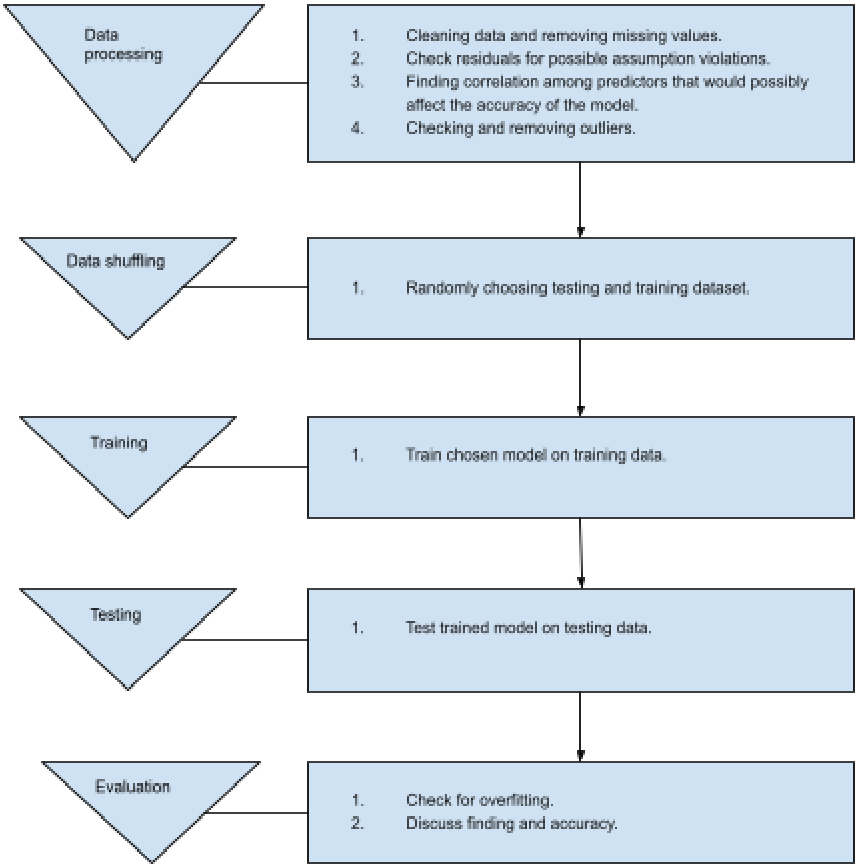
Dataset Background

The dataset utilized in our analysis was retrieved from the Mendeley Diabetes types dataset. The data within the dataset was collected by Iraqi society from the laboratory of Medical City Hospital, the Specialized Center for Endocrinology, and Diabetes-Al-Kindy Teaching Hospital. Data was ethically extracted from patients' files in the Medical City Hospital, the Specialized Center for Endocrinology, and Diabetes-Al-Kindy Teaching Hospital in order to create a diabetes dataset. The diabetes dataset includes 103 (no-diabetes), 53 (pre-diabetic), and 844 (diabetic) patients. Attributes utilized in the diabetes dataset include gender, age, urea, creatine ratio, body mass index, LDL, cholesterol, VLDL, Triglycerides (TG), HDL, HBA1C, and Class. The following descriptions for each of the attributes utilized in the diabetes dataset are provided below.

- Gender: Defines whether a patient is Male or Female.
- Age: The age of a patient in years (min: 20, max: 79).
- Urea: In Mg/dl (min: 0.5, max: 38.9). Urea is a nitrogenous compound that is the end product of the metabolic breakdown of proteins in all mammals.
- Creatinine ratio: µmol/L (min: 48, max: 80). Creatinine is a normal waste product found in urine.
- Body Mass Index: (min: 19, max: 47). Body Mass Index is a measure of an individual's weight-to-height ratio.
- LDL: mmol/L (min: 0.3, max: 9.9). LDL (low-density lipoprotein) is the most commonly found form of cholesterol in the human body.
- Cholesterol: mmol/L (min: 0.0, max: 10.3). Cholesterol is a fat-like substance found in all the cells of the human body.
- VLDL: mmol/L (min: 0.1, max: 35). VLDL (very-low-density lipoprotein) is a cholesterol produced in the liver and released into the bloodstream to supply body tissues with triglycerides.
- Triglycerides (TG): mmol/L (min: 0.3, max: 13.8). Triglycerides are a type of fat found in the bloodstream.
- HDL: mmol/L (min: 0.2, max: 9.9). HDL (high-density lipoprotein) is a cholesterol that absorbs other forms of cholesterol within the body and carries it back to the liver.
- HBA1C: mmol/L (min: 0.9, max: 16). HBA1C (Hemoglobin A1C) defines blood sugar.
- Class: Defines whether a patient is Diabetic, Non-Diabetic, or Predict-Diabetic.

Procedure

Within this paper, different machine learning techniques to analyze our diabetes data set were utilized to make accurate predictions on whether a patient has or could have diabetes. Our main methods for prediction included logistic regression, random forest, and k-nearest neighbor algorithms. Seen below is our analysis process we applied to the dataset:



Data processing

1. Cleaning data and removing missing values.
2. Check residuals for possible assumption violations.
3. Finding correlation among predictors that would possibly affect the accuracy of the model.
4. Checking and removing outliers.

Data shuffling

1. Randomly choosing testing and training dataset.

Training

1. Train chosen model on training data.

Testing

1. Test trained model on testing data.

Evaluation

1. Check for overfitting.
2. Discuss finding and accuracy.

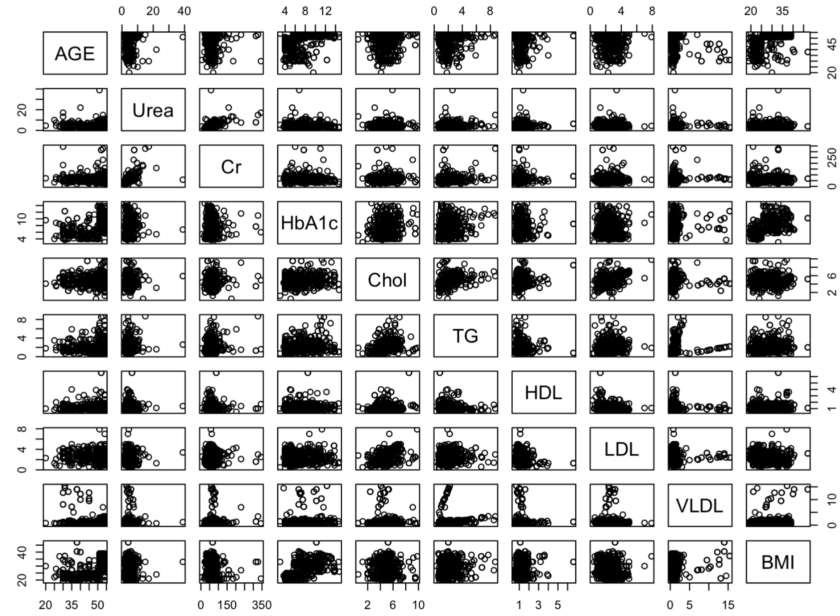
Sequential steps in the proposed system

Formatting the Dataset

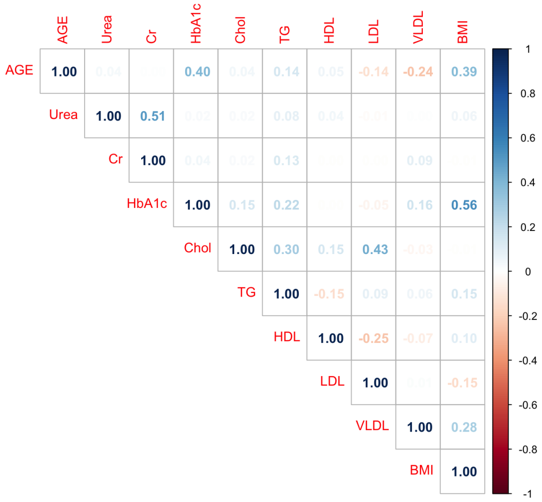
Before beginning our analysis of the dataset, we first had to reorganize the dataset to ease our analysis and to create unskewed results. Within the dataset we found that for patients above the age of 55, the vast majority of them were diagnosed with diabetes which led to a heavy disproportionality in the dataset. To create more of a balance between non-diabetics and diabetics, we did not use these people to create models except for the logistic model.

Correlation Test

Initially, we created a mass scatter plot including all the various variables within our model. This showed us potential correlations within the dataset and provided us with areas to further investigate. One area in particular that we found interesting was TG with VLDL as that scatter plot displayed two differing models.

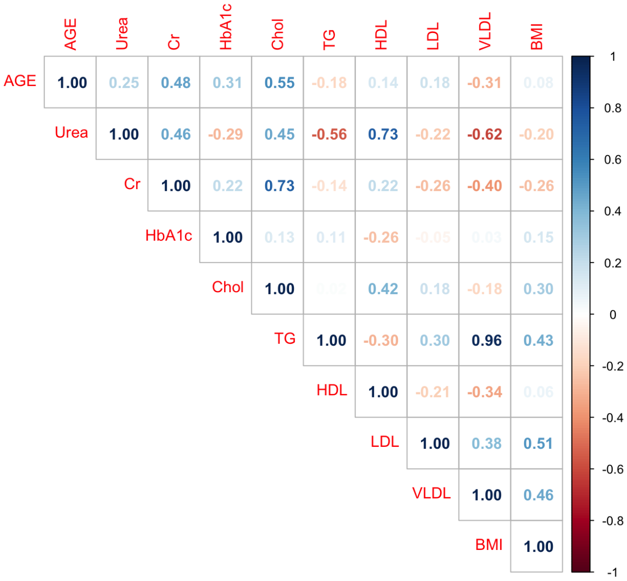


Examining a correlation plot as seen below, we can see that there is little to no correlation between TG and VLDL.



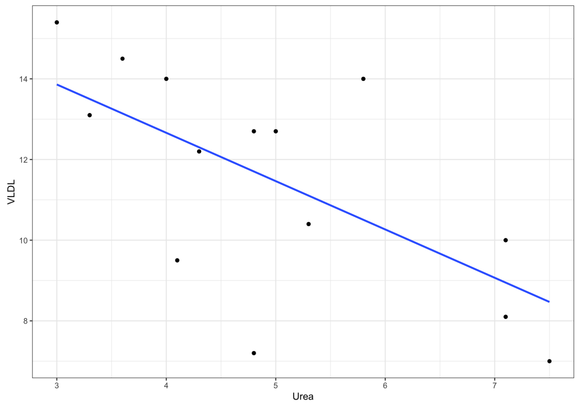
	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
AGE	1.00	0.04	[illegible]	0.40	0.04	0.14	0.05	-0.14	-0.24	0.39
Urea		1.00	0.51	0.02	0.02	0.08	0.04	-0.01	[illegible]	0.06
Cr			1.00	0.04	0.02	0.13	[illegible]	[illegible]	0.09	[illegible]
HbA1c				1.00	0.15	0.22	[illegible]	-0.05	0.16	0.56
Chol					1.00	0.30	0.15	0.43	-0.03	[illegible]
TG						1.00	-0.15	0.09	0.06	0.15
HDL							1.00	-0.25	-0.07	0.10
LDL								1.00	0.01	-0.15
VLDL									1.00	0.28
BMI										1.00

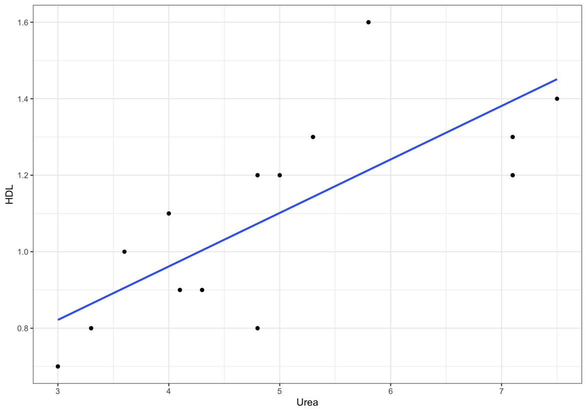
To investigate why two differing models are present within the scatterplot, we split the VLDL values into two: values above 5 and values below 5. Upon doing so, we can see the correlation jumps to an almost perfect correlation of 0.96 when examining values for VLDL of above 5 only as seen below:



	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
AGE	1.00	0.25	0.48	0.31	0.55	-0.18	0.14	0.18	-0.31	0.08
Urea		1.00	0.46	-0.29	0.45	-0.56	0.73	-0.22	-0.62	-0.20
Cr			1.00	0.22	0.73	-0.14	0.22	-0.26	-0.40	-0.26
HbA1c				1.00	0.13	0.11	-0.26	-0.05	0.03	0.15
Chol					1.00	0.02	0.42	0.18	-0.18	0.30
TG						1.00	-0.30	0.30	0.96	0.43
HDL							1.00	-0.21	-0.34	0.06
LDL								1.00	0.38	0.51
VLDL									1.00	0.46
BMI										1.00

The reasoning behind choosing VLDL values of above 5 only is because for values below 5 there was no significant correlation to be found. Additionally, we can see that Urea has a negative correlation with VLDL (left graph), which is supported by our limited background research on the subject. Also, Urea has a positive correlation with HDL (right graph), which is again supported by our research.





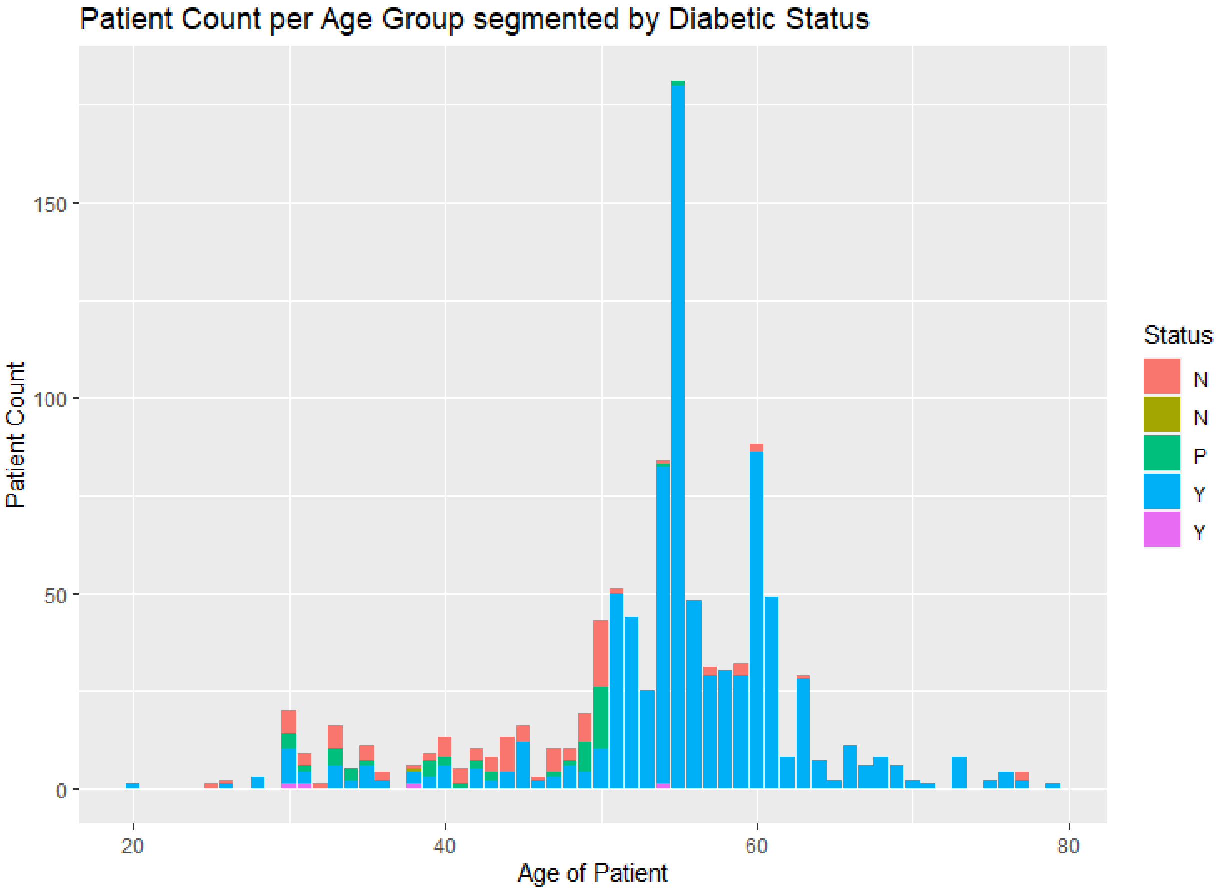
However, certain correlations are only present due to a significant outlier such as the 0.73 correlation between Cr and Chol.

Logistic Regression

Logistic regression is a regression method that uses a transformation of the ordinary least squares linear regression function in order to predict a binomial discrete outcome of 0 or 1. In our case, we would use this model to predict whether people have diabetes or not based on the predictors available in the dataset of patient gender, age, blood sugar (HbA1c), creatine ratio, BMI, LDL cholesterol measure, VLDL cholesterol measure, triglycerides (TG), HDL cholesterol measure, and overall cholesterol measure.

LR Step 1: Cleaning the Data for Logistic Regression

The dataset we used had several noticeable issues. When graphing patient count per age group and displaying the diabetic status of each patient in the bar graph, we see some noticeable issues:



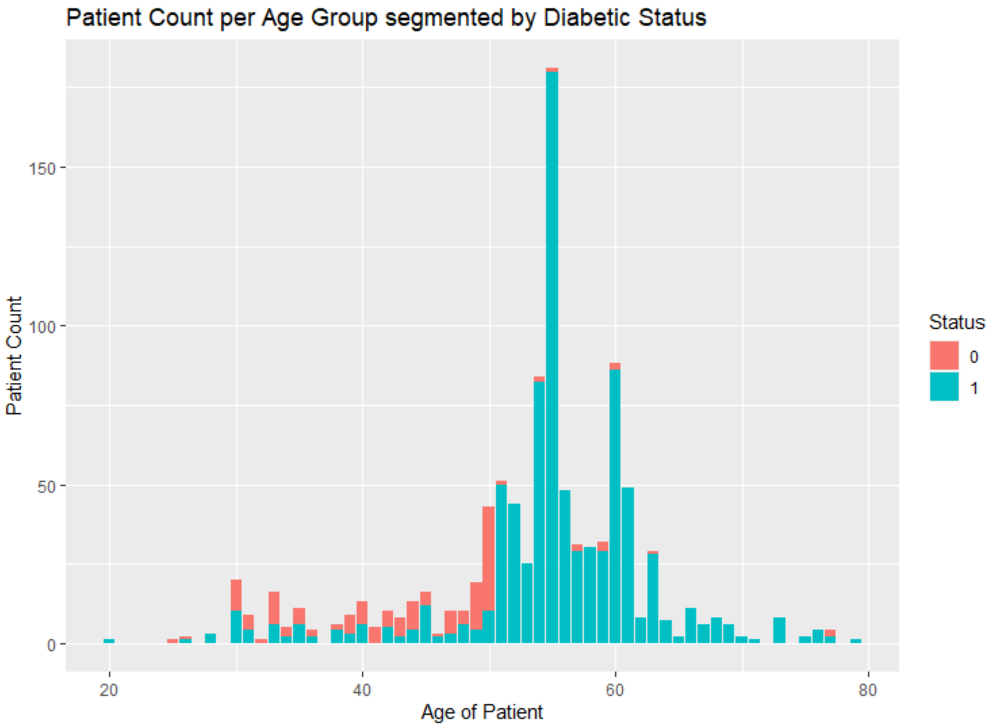
The issues that can be identified are:

- **Mislabelled Data**: there are two categories for N (non-diabetic), and two categories for Y (diabetic).
- **3 Discrete Outcomes**: the dataset lists patients as either being not diabetic, diabetic, or predicted to be diabetic; since logistic regression requires two predictors, people predicted to have diabetes by the data collectors will be excluded (as they constitute only

53/1000 samples); these patients will be used as a validation set at the end of the model-building process in order to determine how similar our predictive model is to the predictive model used to identify these patients as predicted.

- **Imbalanced Data:** above the age of 51, the vast majority of patients are classified as having diabetes, while below the age of 51 there is a much lower proportion of diabetic patients. This will skew the importance of age as a factor in determining diabetes, and can cause other unwanted effects. Within this group of older patients, those of age 55 constitute around 180/1000 observations, and are above 99% diabetic. These patients will not be used in the model building process, but will be used to validate the model.

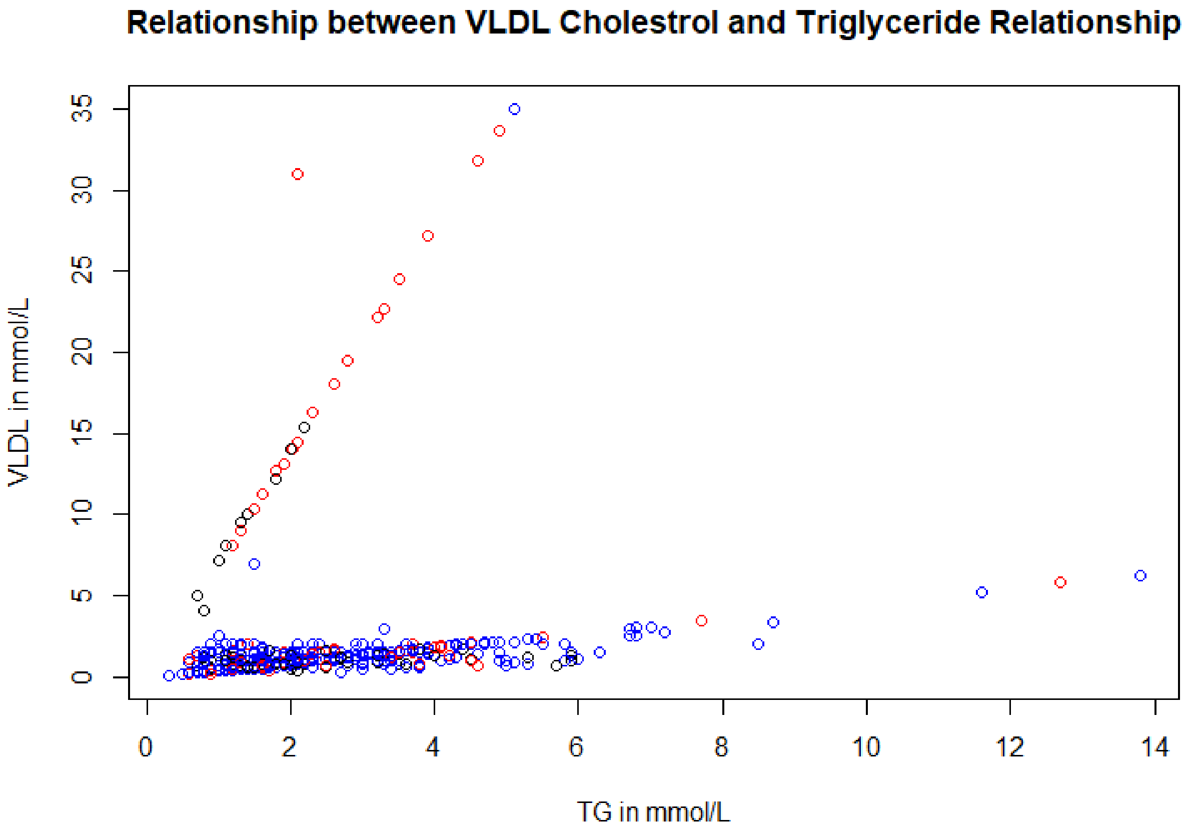
Below is a graph of the dataset with patients predicted to have diabetes extracted and cleaned labeling:



The outcome variables are now 0 and 1 as opposed to “yes” and “no,” which is necessary for logistic regression.

LR Step 2: Removing Individuals of Age 55 from Model-Building

There were a few reasons why people of age 55 were removed from the model building process. One reason is that they were overwhelmingly diabetic compared to younger individuals, and constituted a significant portion of the sample (possibly causing imbalance). The second reason is that their recorded data was unusually correlated, for example:



In red are patients of age 55, and it can be seen that many of them had an unusually higher level of VLDL cholesterol compared to the level of triglycerides in their blood than the rest of the dataset (in blue and black). In fact, 31/60 patients who had abnormally high VLDL levels (above 5) were in this group. For this reason, it is necessary to remove these individuals from the process of model creation, and only use them for validation.

VLDL was also dropped as a possible predictor due to its obvious high correlation with other predictors such as TG (triglyceride levels), as is apparent in this graph.

LR Step 3: Should Older Individuals be included in the Model?

The first pressing question of the model-building process was to determine if older individuals (≥51) should be included at all: was it better to use *only* patients of lower age (<51) to predict the entire sample. The process to determine whether this a better idea or not involved the following process:

- Repeat 50 (n) times:
    - Split the dataset of older + younger patients (excluding patients of age 55) into training and testing using a 40/60 split and stratifying for age.
    - Find the best logistic model for the training set using best subsets regression and AIC as the criteria; calculate the model's AUC on its respective test data, the misclassification rate on the respective test data (the rate of incorrect predictions), and the misclassification rate on the entire model.

- Do the same steps on the dataset of only younger patients (patients of age < 51).

- After 50 loops, calculate the average AUC, average test misclassification rate, and average misclassification rate on the entire dataset. Here are the results:

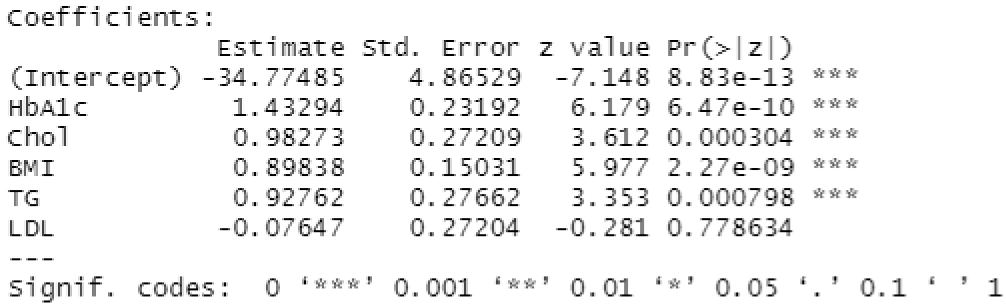
	Model using Older and Younger patients	Model using only Younger patients
Average AUC on Test Sample	0.987	0.915
Average Misclassification Rate on Test Sample	0.038	0.167
Average Misclassification Rate on Entire Sample	0.0265	0.033

When creating models with only younger patients, the models had worse AUCs on their test dataset (meaning a weaker ability to distinguish between true negatives and true positives), and have higher misclassification rates on their test dataset. Therefore, a model using both older and younger patients (excluding those of age 55) will be used.

This method also found a "best model" from the models of older and younger patients. This was done by finding which model had the lowest misclassification rate on its test sample out of all 50 models. The model returned was this one:

- Diabetic Status ~ HbA1c + Chol + TG + LDL + BMI

This model had the following coefficients:



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -34.77485    4.86529  -7.148 8.83e-13 ***
HbA1c         1.43294    0.23192   6.179 6.47e-10 ***
Chol          0.98273    0.27209   3.612 0.000304 ***
BMI           0.89838    0.15031   5.977 2.27e-09 ***
TG            0.92762    0.27662   3.353 0.000798 ***
LDL          -0.07647    0.27204  -0.281 0.778634    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Already, an issue emerges as one of the predictors are not significant. The model has the following statistics:

- On entire dataset:
  - An area under the curve of 0.9944
  - A misclassification rate of 0.232, meaning that the model only falsely predicted values about 2.32% of the time.
- On younger people (age < 51)
  - A misclassification rate of 0.118, meaning that the model only falsely predicted values about 2.32% of the time.
- On older people (age ≥ 51)
  - A misclassification rate of 0%

However, the fact that insignificant predictors made their way into this model shows that this method of creating a “best model” can be inaccurate and results in insignificant information being held over.

LR Step 4: Using K Fold Validation to Pick a Best Model

From the previous step it was determined that data from older and younger patients should be utilized to develop a logistic regression model that best represents patterns in the entire dataset. Now it is time to find the best model for predicting whether a patient has diabetes using the predictors present. This will be done through K-Fold Validation with 10 folds. This process does the following:

- Splits the data into n groups (for this stage we used 10 groups, a standard amount for balancing between specificity (the ability to correctly predict people without diabetes) and sensitivity (the ability to correctly predict people with diabetes).
- Evaluates each of the best models on these 10 groups, and compiles a score of overall error through the cross validation RMSE error statistic.

The resulting 5 best models were the following:

- Diabetic Status ~ HbA1c + Chol + TG + Gender + BMI
- Diabetic Status ~ HbA1c + Chol + TG + BMI
- Diabetic Status ~ HbA1c + Chol + TG + HDL + Gender + BMI
- Diabetic Status ~ HbA1c + Chol + TG + LDL + Gender + BMI
- Diabetic Status ~ Age+ HbA1c + Chol + TG + LDL + Gender + BMI

These are the respective cross-validation errors after running the cross-validation method:

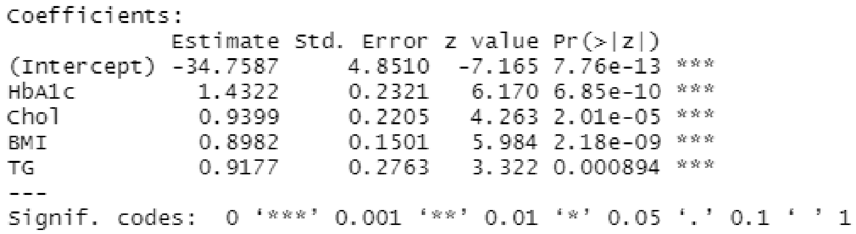
- Model 1: 0.141

- Model 2: 0.137
- Model 3: 0.143
- Model 4: 0.142
- Model 5: 0.144

Because Model 2 (Diabetic Status ~ HbA1c + Chol + TG + BMI) had the lowest cross-validation error score, it will be used as the final model for predicting diabetes. It is worth noting that all models had very similar cross-validation errors.

LG Step 5: Validating the Final Model

Here are the results for evaluating the final model on the entire dataset:



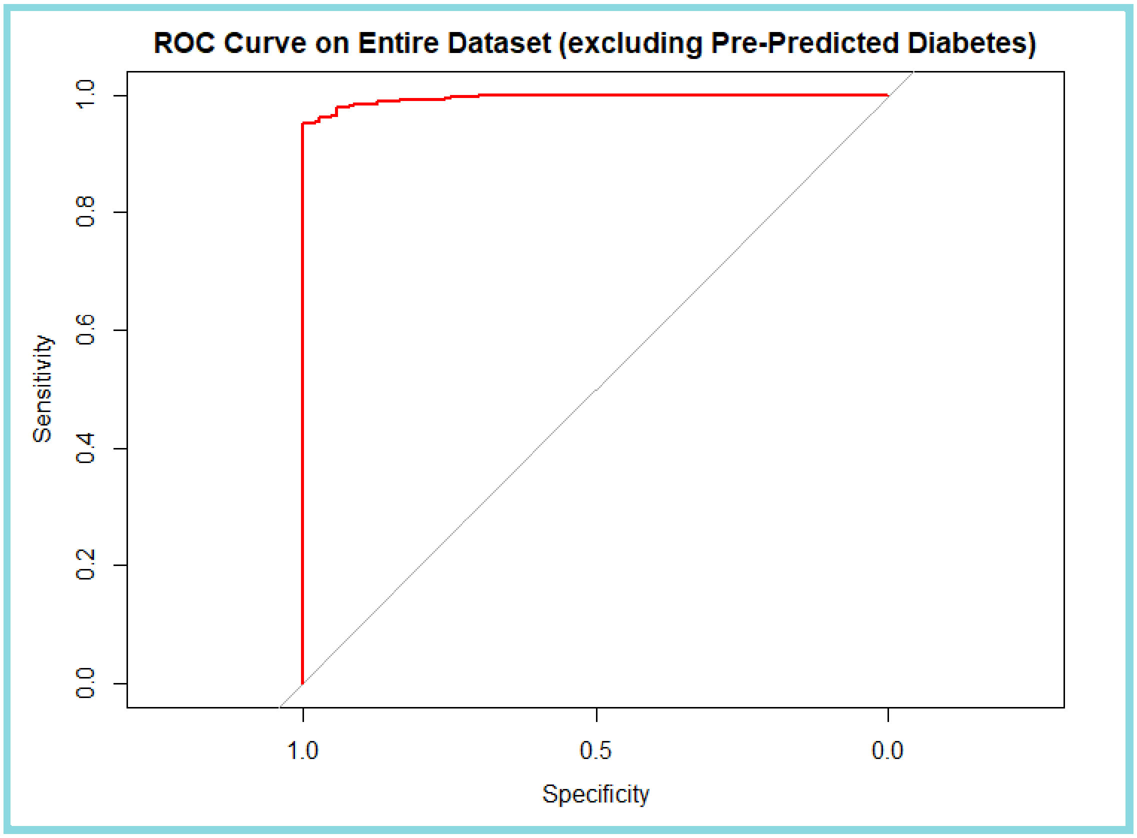
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -34.7587     4.8510  -7.165 7.76e-13 ***
HbA1c         1.4322     0.2321   6.170 6.85e-10 ***
Chol          0.9399     0.2205   4.263 2.01e-05 ***
BMI           0.8982     0.1501   5.984 2.18e-09 ***
TG            0.9177     0.2763   3.322 0.000894 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All coefficients are significant in predicting diabetes. The final function returned is that:

$$form = -34.7587 + 1.4322(HbA1c) + 0.9399(Chol) + 0.8982(BMI) + 0.9177(TG)$$

$$predicted\ odds = \frac{e^{form}}{1+e^{form}}$$

ROC Curve for the Model:



Area Under the Curve: 0.9945 (very high, showing the model can distinguish highly between true positives and true negatives).

Confusion Matrices: Shows the ability of the model to correctly predict people who have diabetes or do not have diabetes. A cutoff predicted odds of having diabetes of 0.5 was used to distinguish between those predicted to have diabetes and those not predicted (a person with a predicted proportion of having diabetes of <0.5 was said to not have diabetes, 0.5 and above was predicted to have diabetes).

Confusion matrix on the entire dataset (excluding the 53 pre-predicted individuals):

	Actually Not Diabetic	**Actually Diabetic**
Predicted Not Diabetic	92	13
Predicted Diabetic	11	831

Misclassification Rate: 0.025

Confusion matrix on patients younger than 51:

	Actually Not Diabetic	**Actually Diabetic**
Predicted Not Diabetic	81	13
Predicted Diabetic	10	83

Misclassification Rate: 0.123

Confusion matrix on patients 51 or older:

	Actually Not Diabetic	**Actually Diabetic**
Predicted Not Diabetic	11	0
Predicted Diabetic	1	748

Misclassification Rate: 0.0013

Confusion matrix on Pre-Predicted Patients: these were patients removed from the initial modeling process due to their status as being pre-predicted to having diabetes. They will be used to see how well this model matches the model used by the data collectors or doctors to determine if people were predicted diabetic.

	Indicated in Dataset as Predicted Diabetic
Not Predicted Diabetic by Model	13
Predicted Diabetic by Model	40

Misclassification Rate: 0.2453, this indicates that there was some difference in how data collectors assigned people as predicted to become diabetic as opposed to how the model we generated did.

LR Step 5: Conclusion

In summary, it appears that one of the better ways to predict a patient having diabetes is by measuring their HbA1c (blood sugar level), their cholesterol level, their triglyceride level (TG), and their BMI (body mass index). This model has slightly higher misclassification errors than the model obtained by looping models obtained from training and test samples (which includes LDL cholesterol levels), but does not retain insignificant predictors. This model resulted in:

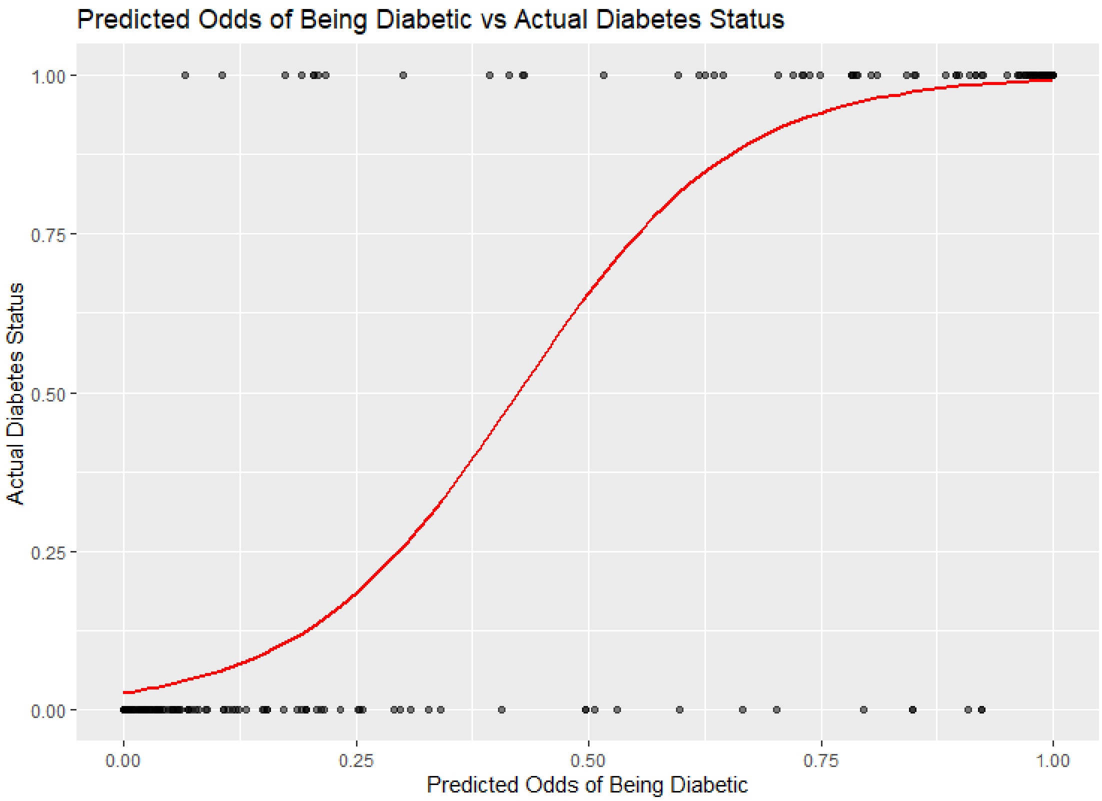
- An ability to predict peoples' diabetic status on all patients, young or older, around 97.5% of the time.
- An ability to predict younger patients' diabetic status about 87.7% of the time.

- An ability to predict older patients' diabetic status almost 100% of the time.

There were several downsides to using a logistic approach, however, including that:

- The imbalance nature of the data resulted in predicted probabilities that were 0 or 1; this is not ideal as the logistic functions' predicted odds should always be between 0 and 1, as the logistic curve should ideally be asymptotic to these values.
- The model was much at predicting diabetes in younger individuals, possibly also due to the imbalance in the dataset.

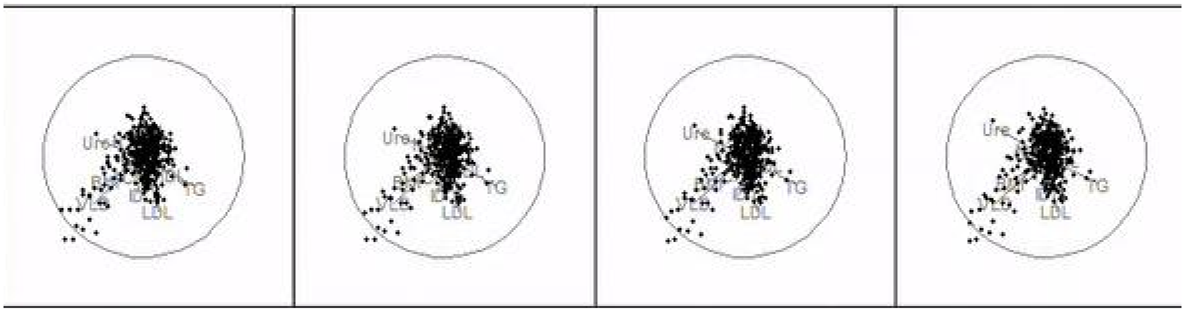
LR Step 6: Logistic Prediction Curve:



The curve of the model's predicted odds of being diabetic matches closely to how the logistic curve should look, another piece of evidence relating to the model's goodness of fit on the data.

K Nearest Neighbor

In hopes of predicting a patient's diabetes status we used the k nearest neighbor algorithm. From the animation below we can see some slightly clustered data, but it is not obvious.



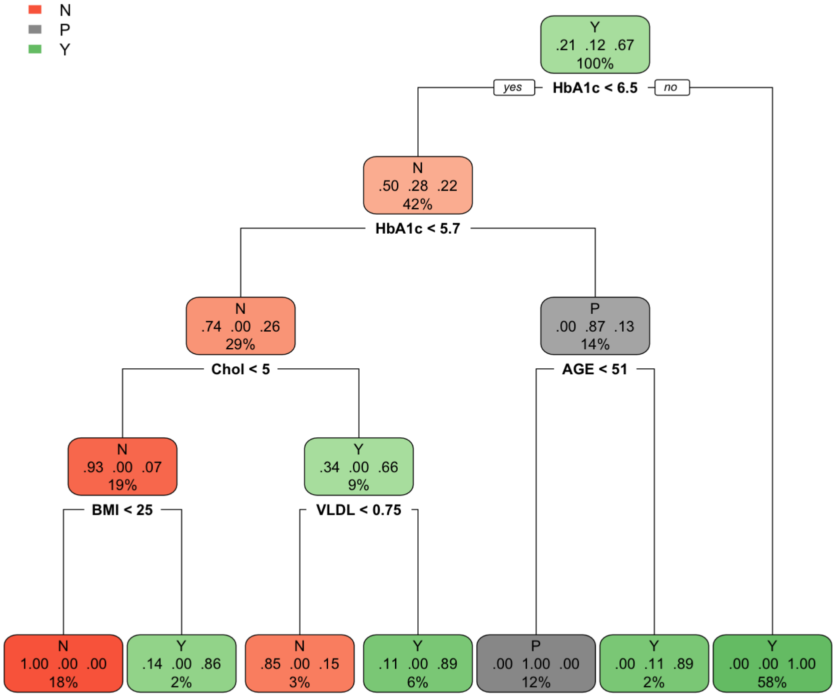
Our expectation would be that the k nearest neighbor algorithm would not be very successful. Before running the algorithm we would normalize data using the following formula:

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

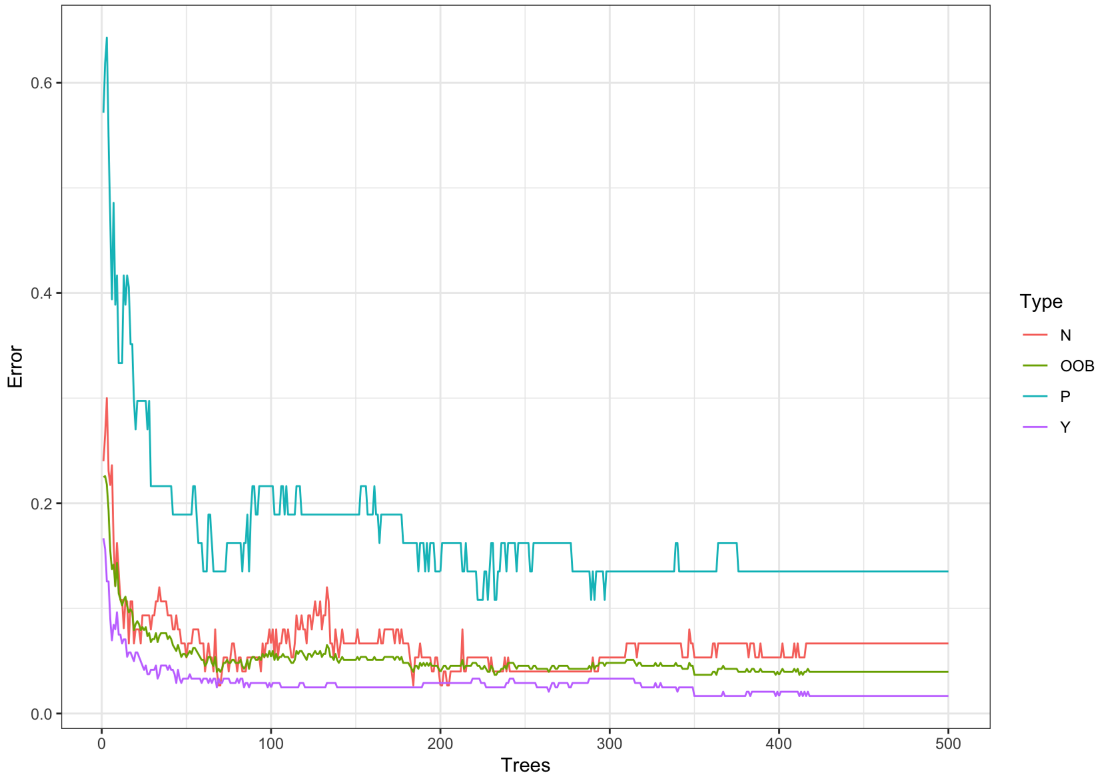
For testing and training purposes we would randomly split the dataset to 80% training and 20% testing. To run the algorithm we used four clusters since it proved to have the lowest error when compared to other variants. We found our accuracy to be 86.67%, which is expected since our data is not as clustered.

Classification Tree and Random Forest

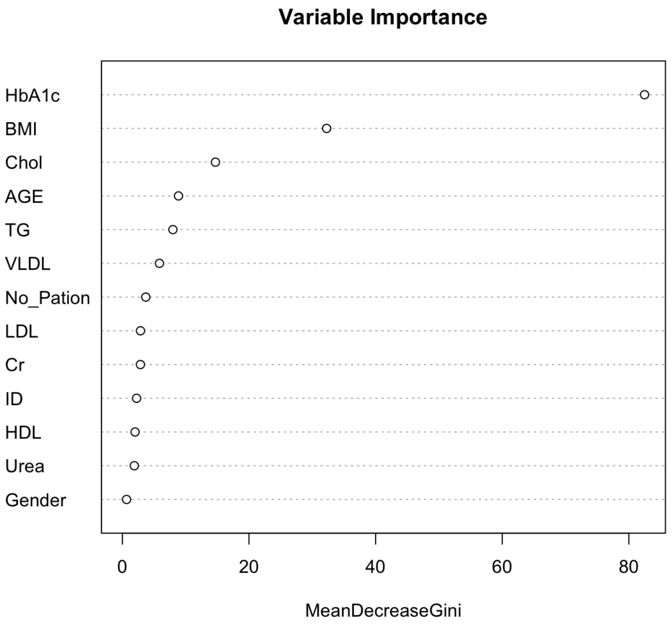
Shown below is the classification tree which makes a decision according to the threshold value, using HbA1c, Chol, BMI, VLDL, and Age as predictors. According to the tree, patients that have Hba1c higher than 6.5 are predicted to have diabetes. As seen by the classification tree below, it gives high priority to the HbA1c level.



Decision trees are a great way of visualizing predictors that are important, but, because it is very sensitive to training data, using decision trees for predictions could result in high variance. Alternatively, we decided to use the random forest algorithm for prediction. Since random forest uses multiple decision trees, randomly selecting predictors at each step, it makes this algorithm great to use for training and testing datasets. Its randomized feature makes a random forest much more accurate than a decision tree. A number of trees for the random forest was selected according to the graph below:



At 500 trees, error rates have been minimized leading us to set 500 trees as the desired value to be used for training and testing the dataset. After additional testing, the number of variables randomly sampled as candidates at each split was chosen to be six, at which the error rate is minimized. Creating a random tree with the specified arguments, we can see that the most important variables within our model are HbA1c, BMI, and Chol.



For testing purposes we would randomly divide data, 80% for training and 20% for testing. As a result, we got a 0.022 percent error rate.

Evaluation

Having initially created a mass scatter plot that included all the various variables within our model, we noticed an unusual plot between TG and VLDL, displaying two differing models. To investigate, we created a correlation matrix which did not give us any indications. Using a different approach, we split VLDL into two levels: above 5 and below 5. These two levels were chosen since that was the splitting point between the two models. Creating a correlation matrix consisting of VLDL values above 5 would depict a positive 0.96 percent correlation between TG and VLDL. Additionally, two more interesting indications became present which were a -0.62 correlation between Urea and VLDL and a 0.73 correlation between Urea and HDL. These findings aligned accurately with our background research. Since VLDL particles often carry TG (triglycerides) within the bloodstream, high VLDL levels correlate with high TG levels.

Concerning the relationship between Urea and VLDL, as higher levels of Urea (waste product) are removed from the body, VLDL (a ‘bad’ cholesterol) decreases. With the relationship between Urea and HDL, as HDL (a ‘good’ cholesterol that carries ‘bad’ cholesterol to the liver) increases, the amount of Urea (waste product) in a patient’s urine also increases.

Having ran a logistic regression and used K-fold validation to pick the best logistic model ($form = -34.7587 + 1.4322(HbA1c) + 0.9399(Chol) + 0.8982(BMI) + 0.9177(TG)$), we found that one of the better ways to predict a patient having diabetes is by measuring their HBA1C (blood sugar level), their cholesterol level, their TG (triglyceride) level, and their BMI (body mass index). The chosen logistical model resulted in:

- 97.5% accuracy in predicting the diabetic status of all patients (below 50 years old and above 50 years old).
- 87.7% accuracy in predicting the diabetic status of patients below 50 years old.
- 100% accuracy in predicting the diabetic status of patients above 50 years old.

We found that the K-nearest neighbor algorithm was not very capable in predicting the diabetes status of a patient effectively, as evident by an accuracy rating of 84.44 percent as well as by the apparent loose clustering. This can be seen through an animation of the K-nearest neighbor plots where each of the graphs showed loose clustered data, furthering our reasoning behind assuming that the K-nearest neighbor algorithm would not perform effectively on our chosen dataset.

When using Random Forest, we found that setting the value of the training dataset and the testing dataset to 500 trees would lead to an optimal minimization of error rate. The resultant error rate of our chosen model was 0.022 percent, suggesting that we would be around 98 percent accurate in predicting diabetes among patients in the dataset with this chosen model.

Citations

Center for Disease Control and Prevention. (2021, December 16). What is diabetes? Retrieved May 9, 2022, from https://www.cdc.gov/diabetes/basics/diabetes.html

Mayo Clinic Staff. (2021, January 20). Type 2 diabetes. Retrieved May 9, 2022, from https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., . . . Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice, 157*, 107843. doi:10.1016/j.diabres.2019.107843

Sharma, A. (2020, May 12). Decision Tree vs. Random Forest - which algorithm should you use? Retrieved May 9, 2022, from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/