DIABETES

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

INTRODUCTION

- Diabetes affects over 450 million people
- One in two people with diabetes don’t know they have diabetes
- Many serious health complications
- We chose to predict possibility of having diabetes

DATASET USED

- Dataset obtained from the cloud-based repository, Mendeley Data.
- Dataset consists of the medical information of 1,000 Iraqi patients from two Iraqi hospitals.
- Attributes in the dataset pertain to the tendency of diabetes in a patient such as:
  - Age
  - Gender
  - Creatinine Ratio
  - HBA1C (Hemoglobin A1C)

MENDELEY

PROCEDURE

01 DATA PROCESSING

- Cleaning Data
- Checking Residuals
- Finding Correlations

02 DATA SHUFFLING

- Randomly choosing testing and training dataset

03 TRAINING

- Train a chosen model on training dataset

04 TESTING

- Test trained model testing dataset

05 EVALUATION

- Check for overfitting
- Discuss findings and accuracy of results

METHODS OF ANALYSIS

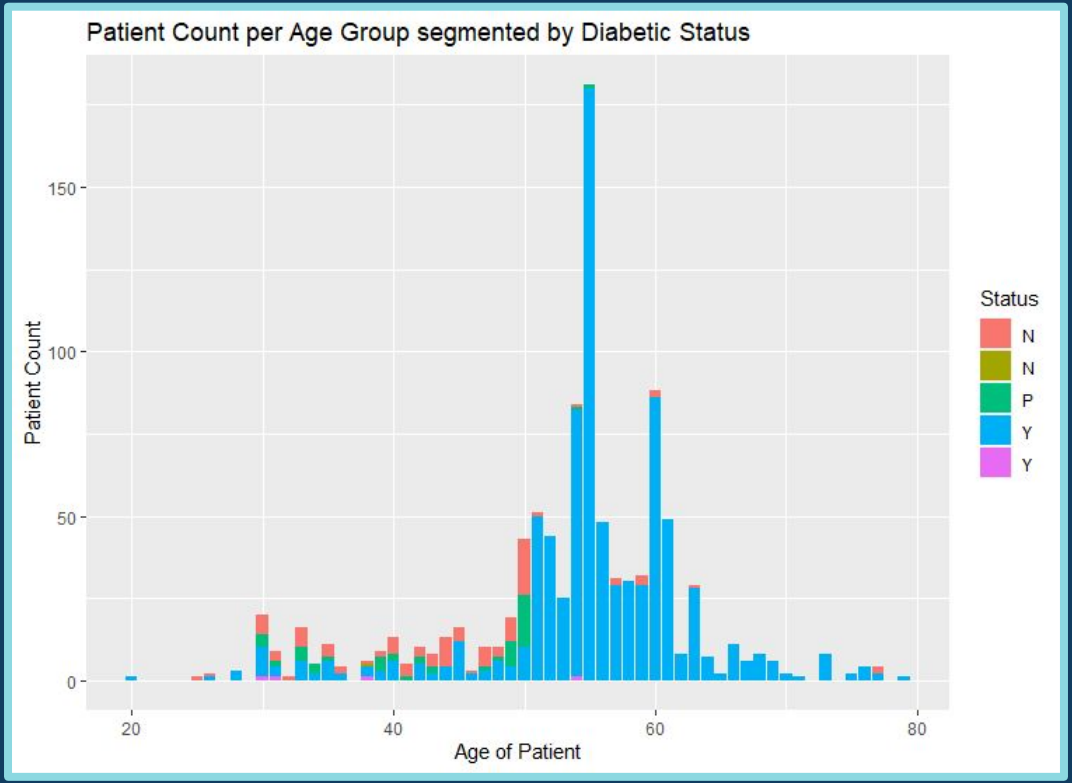
- CORRELATION TEST
- LOGISTIC REGRESSION
- RESIDUAL ANALYSIS
- K-NEAREST NEIGHBOR
- CLASSIFICATION TREE/RANDOM FOREST

DATA PREPARATION

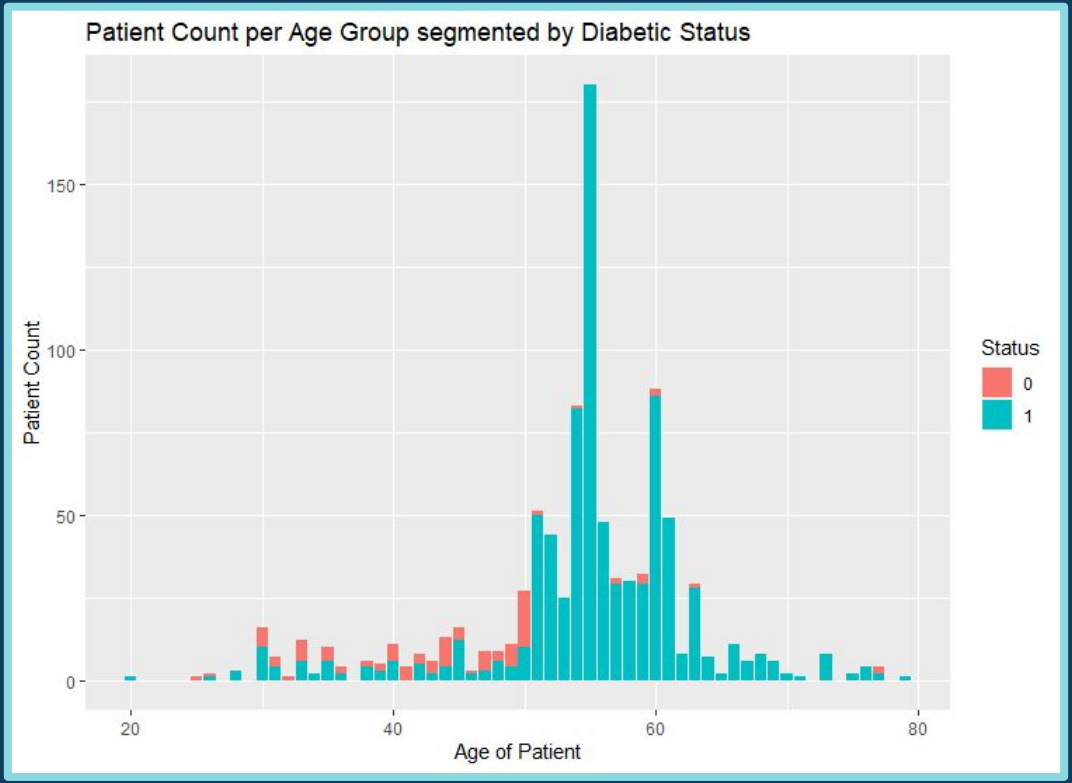
KEY ISSUES:

- 3 default outcomes, need to be reduced to 2
  - Y (has diabetes)
  - N (does not have diabetes)
  - P (predicted to have diabetes)
- Predicted patients were labelled as not having diabetes
- Mislabelled Data
- Skewed Response

MISLABELLED AND IMBALANCED DATA

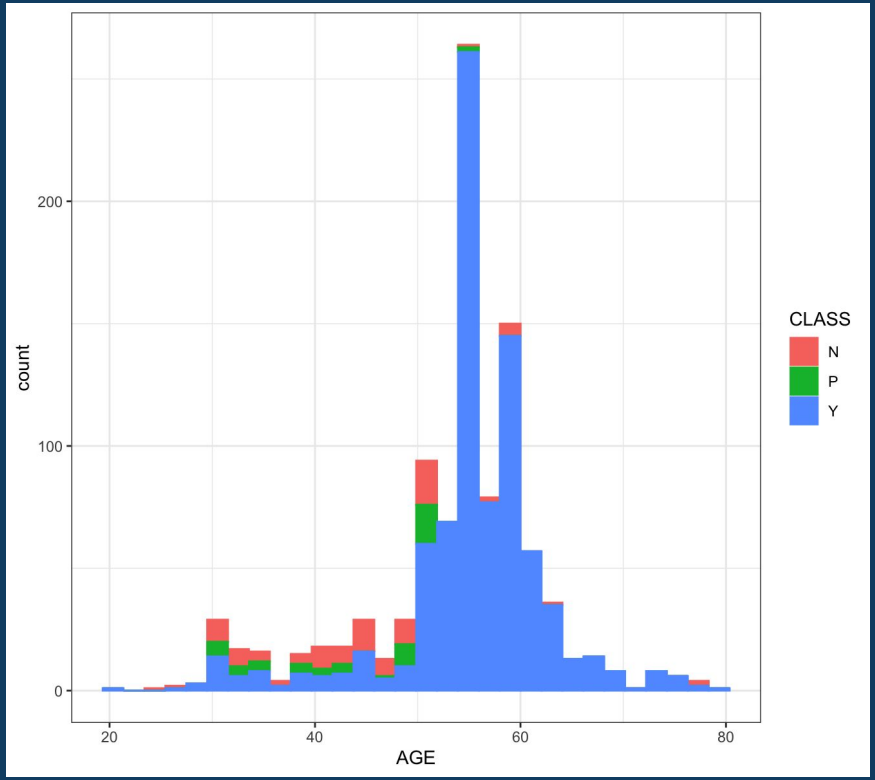


MISLABELLED AND IMBALANCED DATA

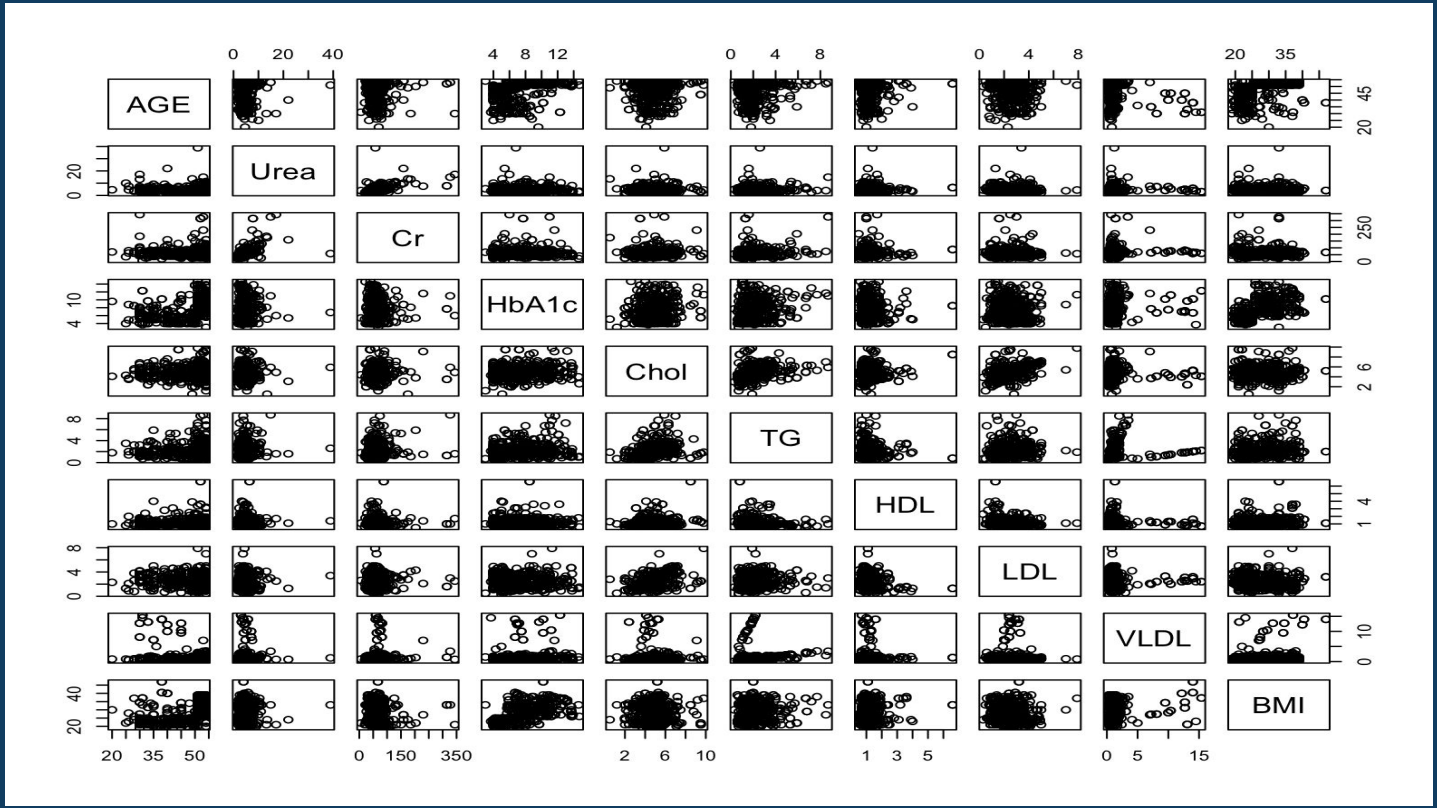


EXCLUDING DATA POINTS

- We removed all patients aged 55 and above for random forest
- Too frequent diabetes in this age group so it hurt our prediction capabilities
- Leaving them out let us predict more accurately for the random forest

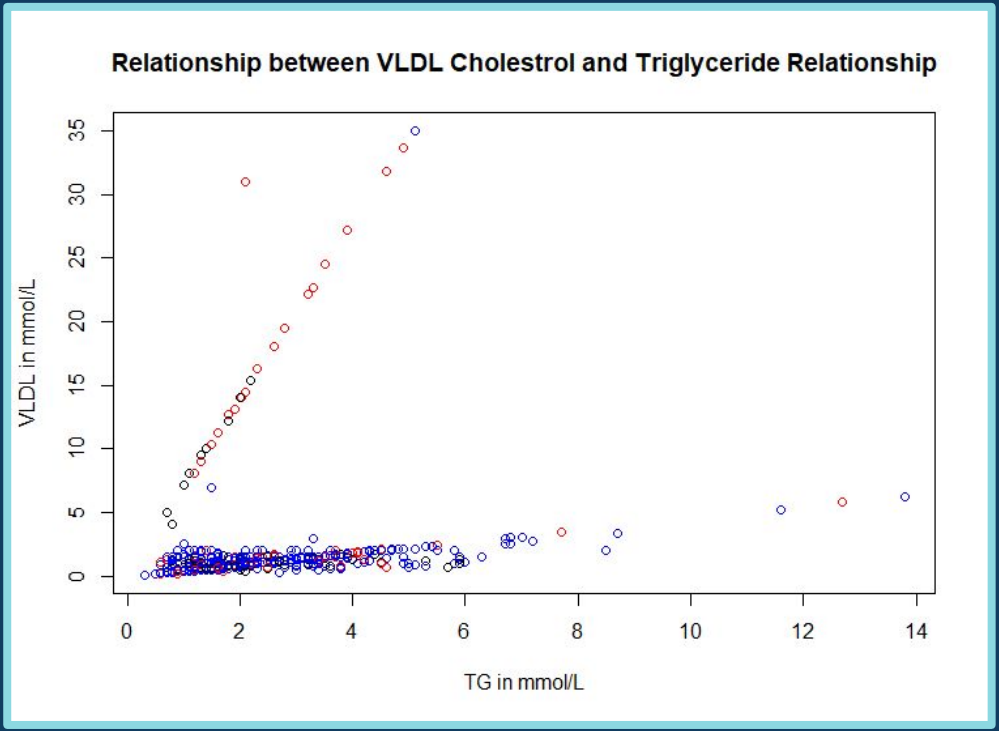


PAIRS SCATTER PLOT

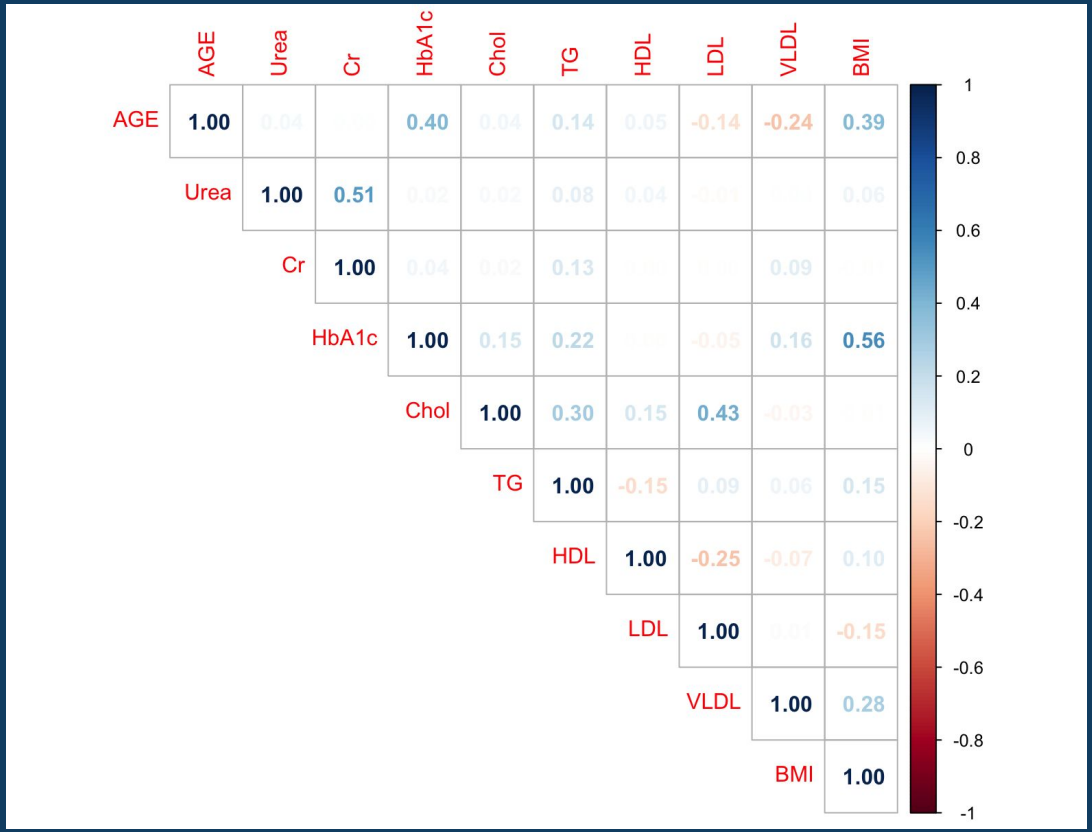


UNUSUAL DATA

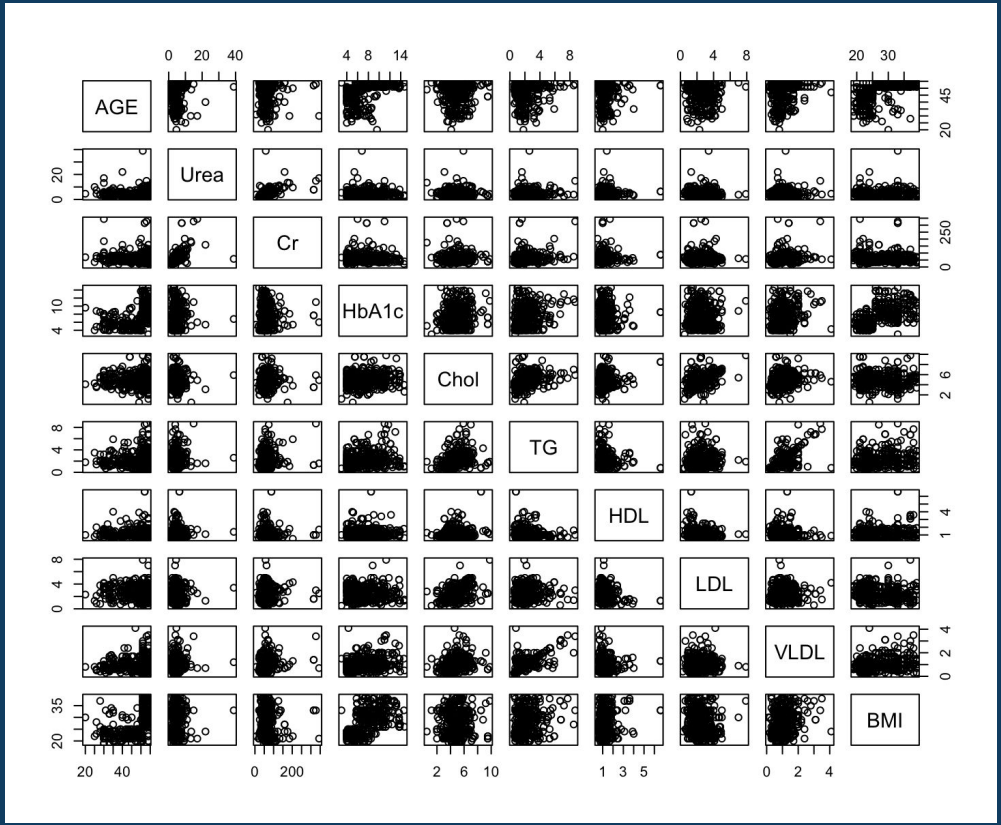
- 60 people in the sample had an unusual relationship between the presence of VLDL and TG



CORRELATION MATRIX

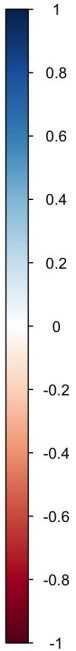


SCATTER PLOT VDVL < 5

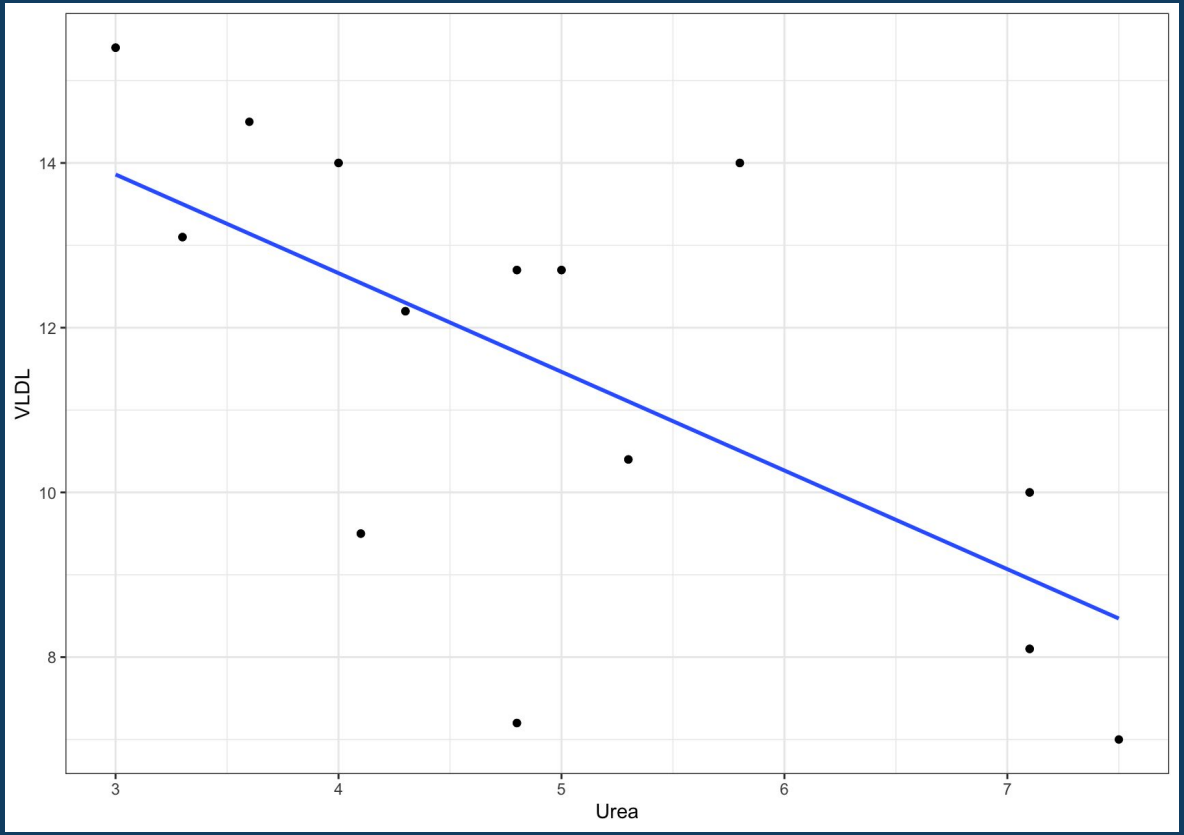


CORRELATION VLDL > 5

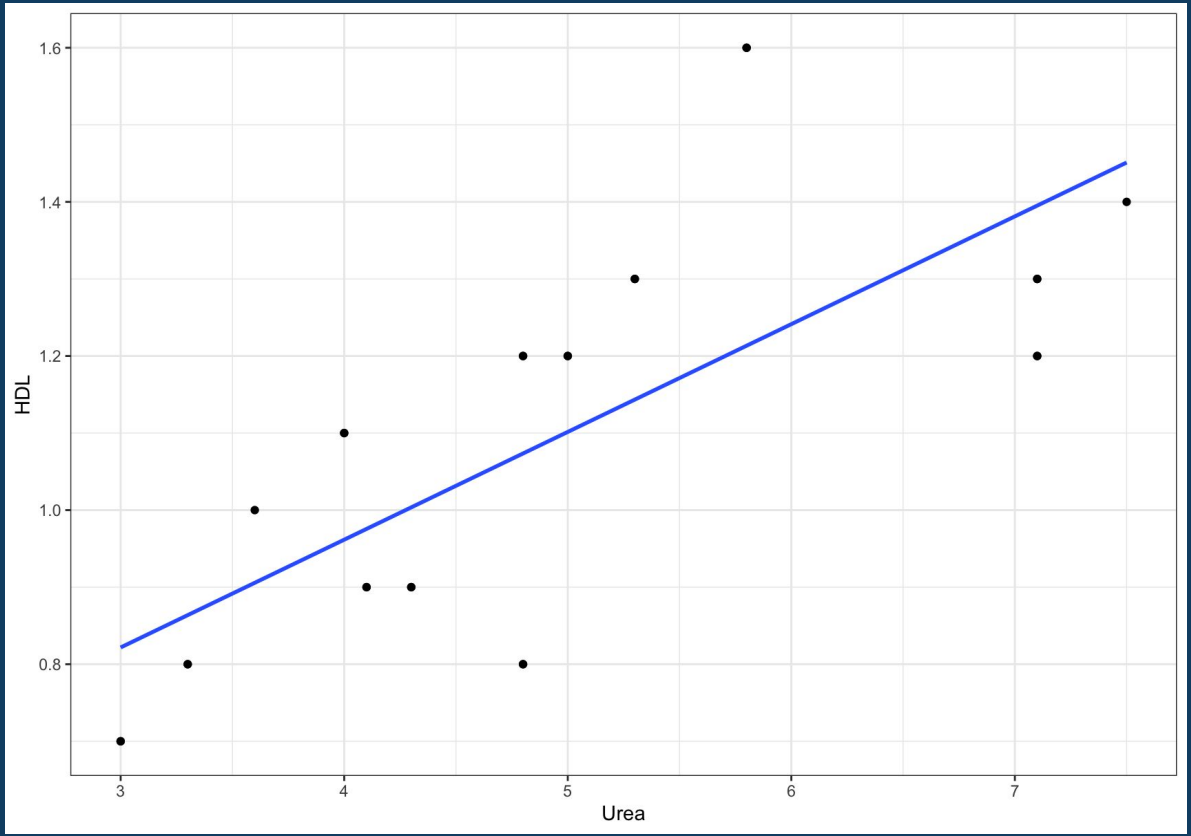
	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
AGE	1.00	0.25	0.48	0.31	0.55	-0.18	0.14	0.18	-0.31	0.08
Urea		1.00	0.46	-0.29	0.45	-0.56	0.73	-0.22	-0.62	-0.20
Cr			1.00	0.22	0.73	-0.14	0.22	-0.26	-0.40	-0.26
HbA1c				1.00	0.13	0.11	-0.26	-0.05	0.03	0.15
Chol					1.00	0.02	0.42	0.18	-0.18	0.30
TG						1.00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



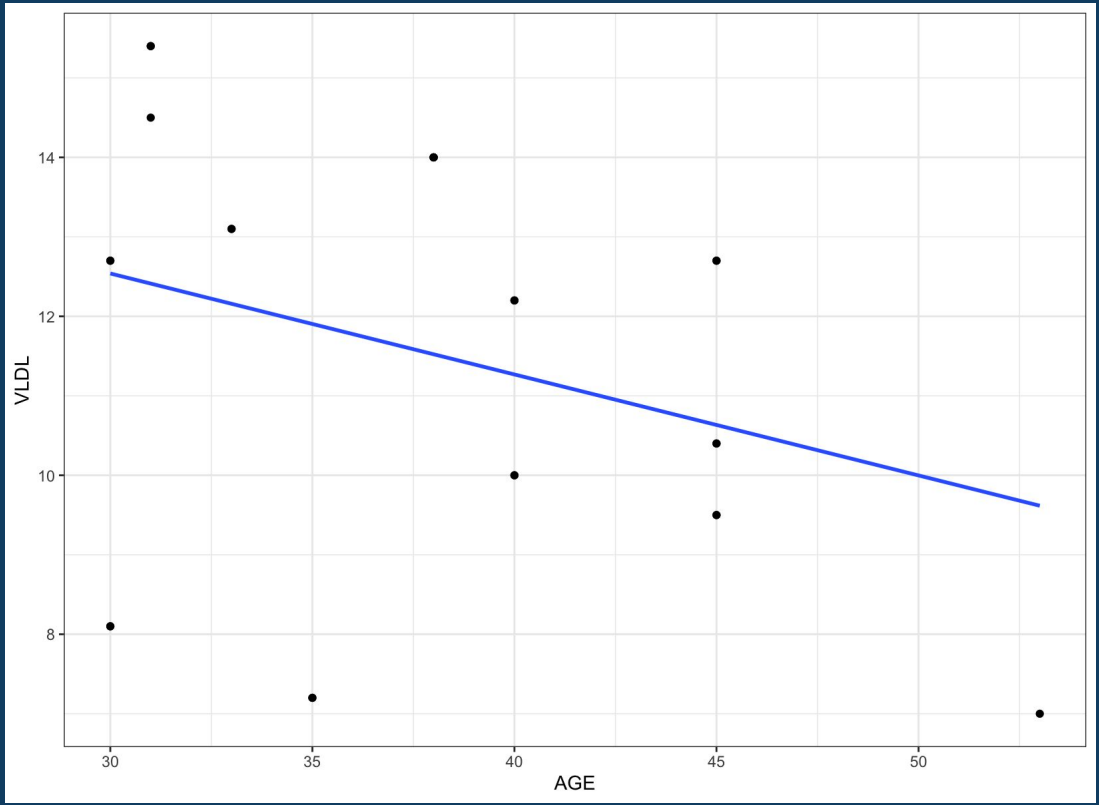
VLDL BY UREA



HDL BY UREA



VLDL BY AGE

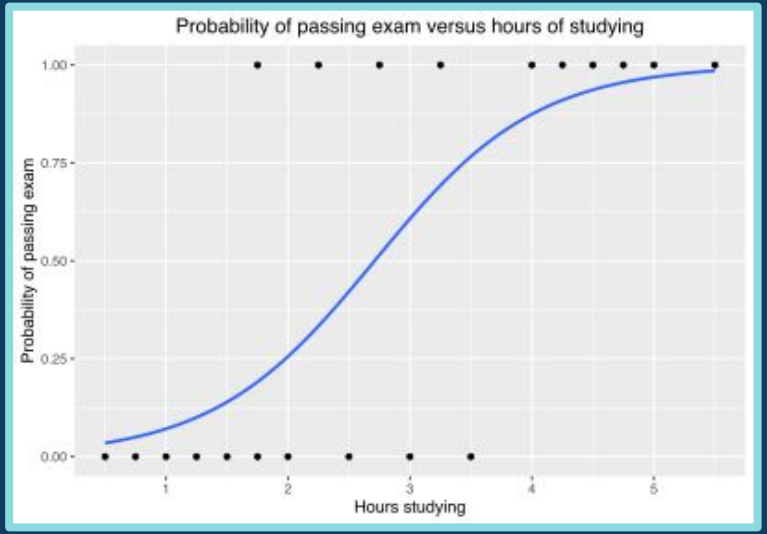


LOGISTIC REGRESSION

What is logistic regression?

- Prediction of a binomial discrete outcome (in this case, having diabetes or not having diabetes)
- Uses a transformation of the least squares linear regression formula to obtain predictions

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$



MODEL BUILDING

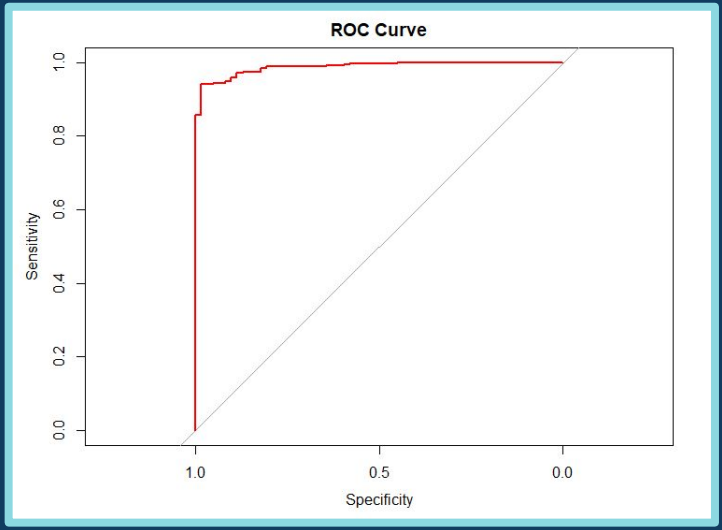
Who should be included in model building?

- Old people and young people or just young people?
  - Usage of Stratified Sampling followed by Best Subsets Regression
  - Build these models first with old and young people (not of age 55), and then just young people
- 40/60 training-test split, stratified based on age

```
stratified_sample <- diabetes_without55 %>%
  group_by(CLASS) %>%
  mutate(num_rows=n()) %>%
  sample_frac(0.4, weight=num_rows)
test  <- anti_join(diabetes_without55, stratified_sample, by = 'row_num')
```

BSR USING YOUNG + OLD PATIENTS

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -51.13099   16.37375  -3.123  0.00179 **
Cr            0.03523    0.02176   1.619  0.10541
HbA1c         1.53075    0.46641   3.282  0.00103 **
Chol          0.88844    0.36198   2.454  0.01411 *
TG            1.32823    0.52473   2.531  0.01137 *
BMI           1.45540    0.52569   2.769  0.00563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Area under the curve: 0.9878

Confusion Matrix (testing split)

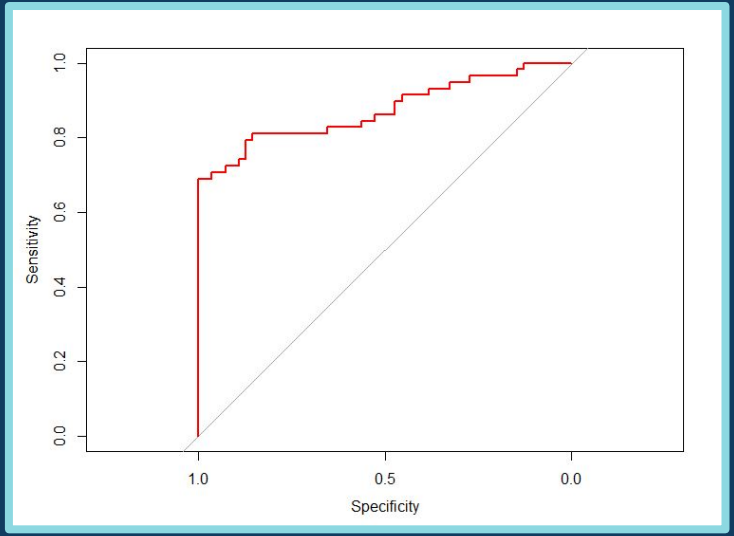
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  51  11
         1  11 387

               Accuracy : 0.9522
```

Confusion Matrix (all data)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  89  14
         1  15 829

               Accuracy : 0.9694
```

BSR USING YOUNG PATIENTS ONLY

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -31.5142     9.3015  -3.388 0.000704 ***
Urea         -0.1999     0.1613  -1.239 0.215292
HbA1c         1.9606     0.5629   3.483 0.000495 ***
TG            0.5958     0.3877   1.537 0.124354
BMI           0.9072     0.3229   2.809 0.004963 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
Area under the curve: 0.8734
```

Confusion Matrix (testing split)

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 48  7
         1 14 44

               Accuracy : 0.8142
```

Confusion Matrix (all data)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  93  10
         1  21 823

               Accuracy : 0.9673
```

MODEL BUILDING

Set with older and younger people had better predictive capabilities on its test sample and it was a better choice. Both models return relatively low misclassification rates on the entire dataset.

- Best model returned after looping the best subset method 50 times resulted in a misclassification rate of 0.0201 on the test dataset:
  - **Class ~ HbA1c* + Chol*+ TG + LDL +GenderM + BMI* *=p value less than 0.05**

Older + Younger Sample:

- Average AUC of 0.987 on test sample
- Average Misc. Rate of 0.040 on test sample
- Average Misc. Rate of 0.029 on full dataset

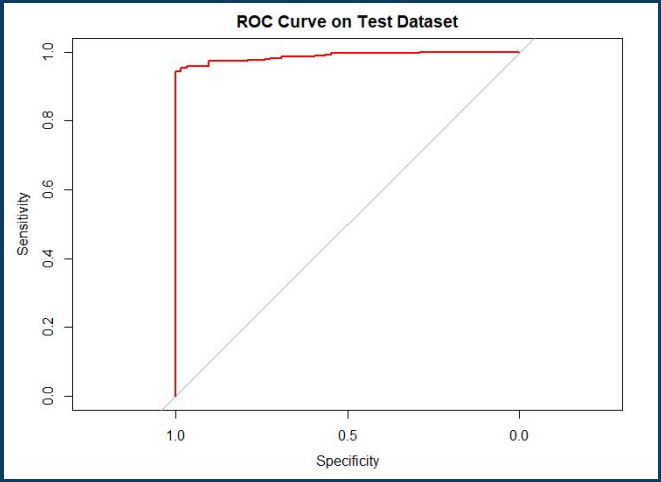
Younger Sample:

- Average AUC of 0.913 on test sample
- Average Misc. Rate of 0.162 on test sample
- Average Misc. Rate of 0.034 on full dataset

VALIDATION ON MODEL

Model Chosen: HbA1c + Chol + BMI

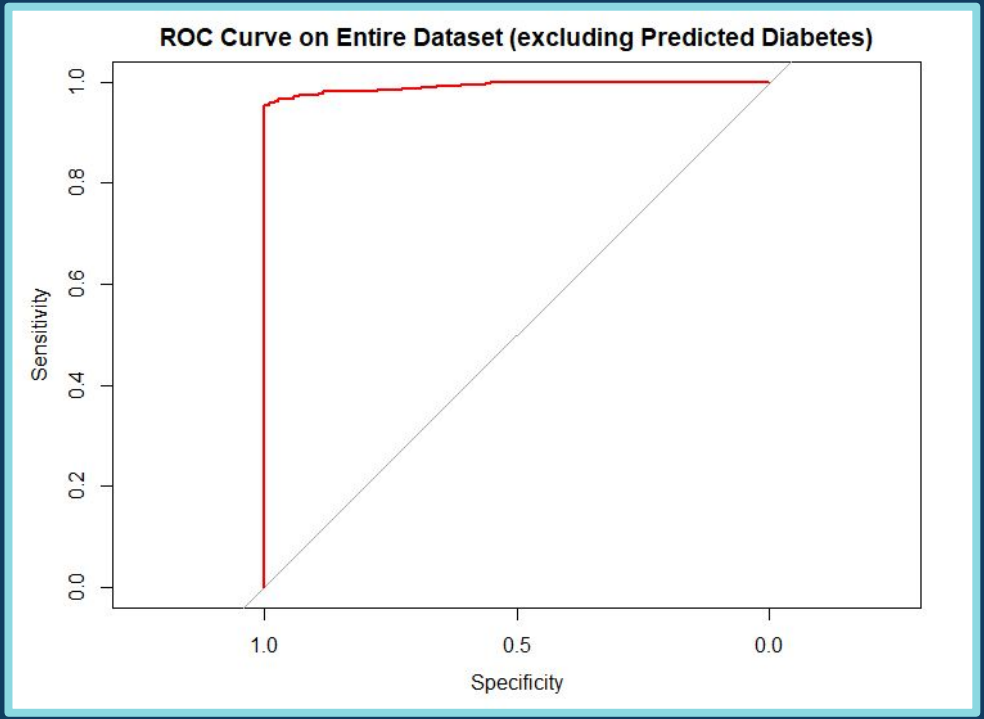
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -30.6343     7.4827  -4.094 4.24e-05 ***
HbA1c         1.3372     0.3420   3.910 9.22e-05 ***
Chol          0.5510     0.2669   2.065  0.03896 *
BMI           0.8840     0.2696   3.280  0.00104 **
---
```



	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	57	17
Predicted Diabetic	5	381

Misclassification Rate of 0.0471

AUC




```
Area under the curve: 0.9913
```

CONFUSION MATRICES

Entire Dataset aside from
Pre-Predicted Individuals

	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	93	23
Predicted Diabetic	10	821

Misclassification Rate of
0.0348

CONFUSION MATRICES

Young Individuals (<51)

	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	81	23
Predicted Diabetic	10	73

Misclassification Rate of
0.144

CONFUSION MATRICES

Older Individuals (≥51)

	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	12	0
Predicted Diabetic	0	748

Misclassification Rate of
0%

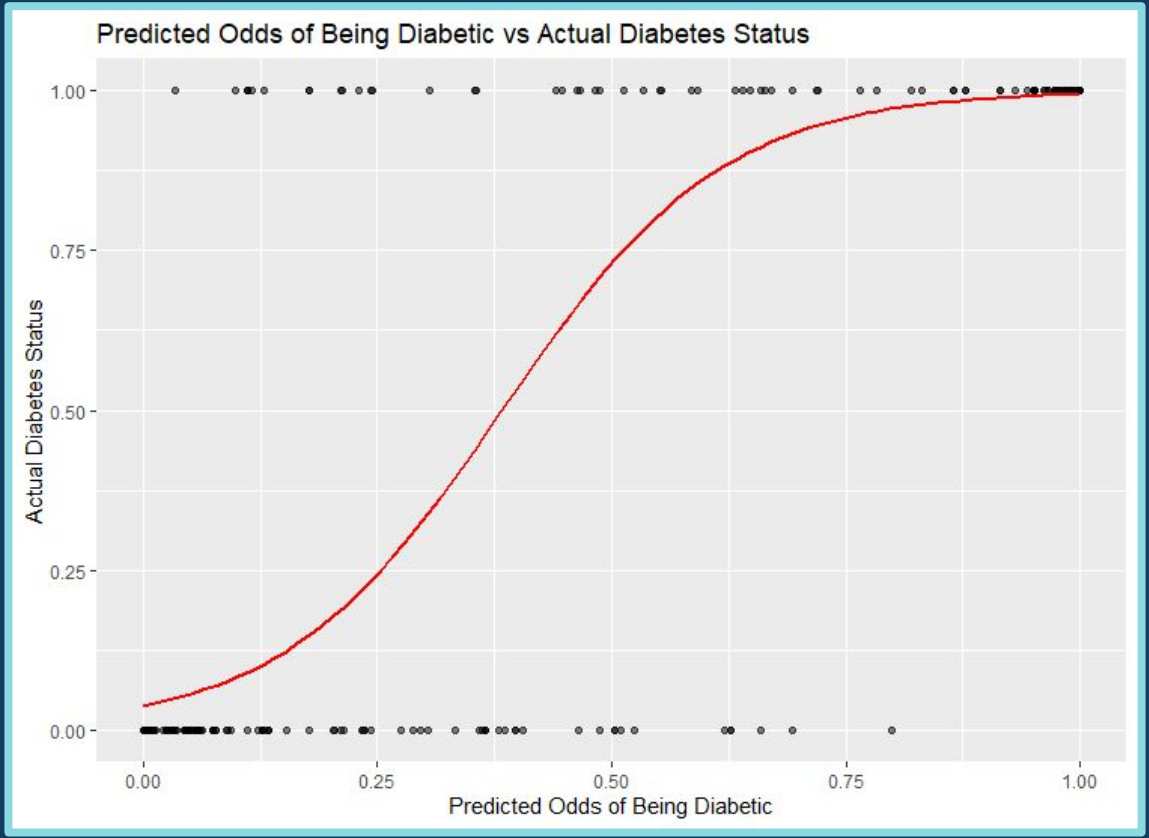
CONFUSION MATRICES

Pre-Predicted Individuals

	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	0	13
Predicted Diabetic	0	40

Misclassification Rate of 0.245

GRAPH


Predicted Odds of Being Diabetic vs Actual Diabetes Status
Actual Diabetes Status
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Predicted Odds of Being Diabetic

K FOLDS CROSS VALIDATION

Process:

- Generate 5 best models using best subsets and full dataset (without predicted)
- Split dataset into 10 folds, evaluate all models on each fold
- Return model with lowest cross-validation errors

```
best.logmodel <-
  glmulti(CLASS ~ AGE + Urea + Cr + HbA1c + Chol + TG + HDL + LDL + Gender + BMI, data = diabetes_edit,
          level = 1,                # No interaction considered
          method = "h",             # Exhaustive approach
          crit = "aic",             # AIC as criteria
          confsetsize = 5,          # Keep 5 best models
          plotty = F, report = F,   # No plot or interim reports
          fitfunction = "glm",      # glm function
          family = binomial)        # binomial family for logistic regression

print(best.logmodel@formulas[1]) #model 1
print(best.logmodel@formulas[2]) #model 2
print(best.logmodel@formulas[3]) #model 3
print(best.logmodel@formulas[4]) #model 4
print(best.logmodel@formulas[5]) #model 5
```

5 BEST MODELS AND ERRORS

5 Best Models:

- Diabetes ~ HbA1c + Chol + TG + Gender + BMI
- Diabetes ~ HbA1c + Chol + TG + BMI
- Diabetes ~ HbA1c + Chol + TG + HDL + Gender + BMI
- Diabetes ~ HbA1c + Chol + TG + LDL + Gender + BMI
- Diabetes ~ Age + HbA1c + Chol + TG + LDL + Gender + BMI

Respective Results:

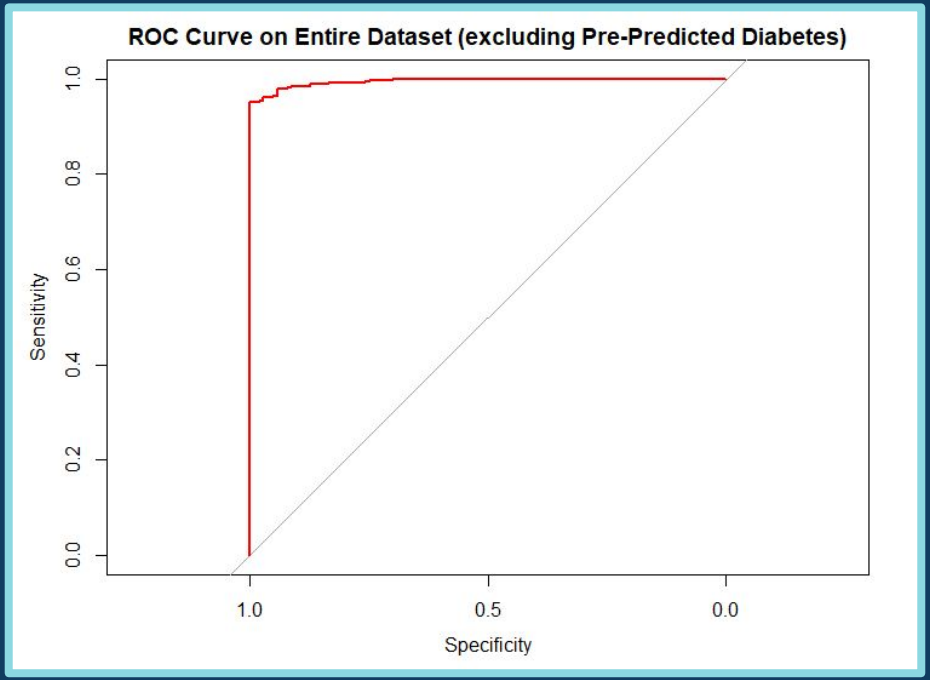
- **Model 1**: 0.1412 **Model 2**: 0.1372 **Model 3**: 0.1433
- **Model 4:** 0.1421 **Model 5**: 0.1437

MODEL 2

Diabetes ~ HbA1c + Chol + TG + BMI

```
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
```

AUC




```
Area under the curve: 0.9945
```

CONFUSION MATRICES

Entire Dataset aside from
Pre-Predicted Individuals

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

Misclassification Rate of
0.025

CONFUSION MATRICES

Younger Individuals (<51)

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

Misclassification Rate of 0.123

CONFUSION MATRICES

Older Individuals (≥51)

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

Misclassification Rate of 0.013

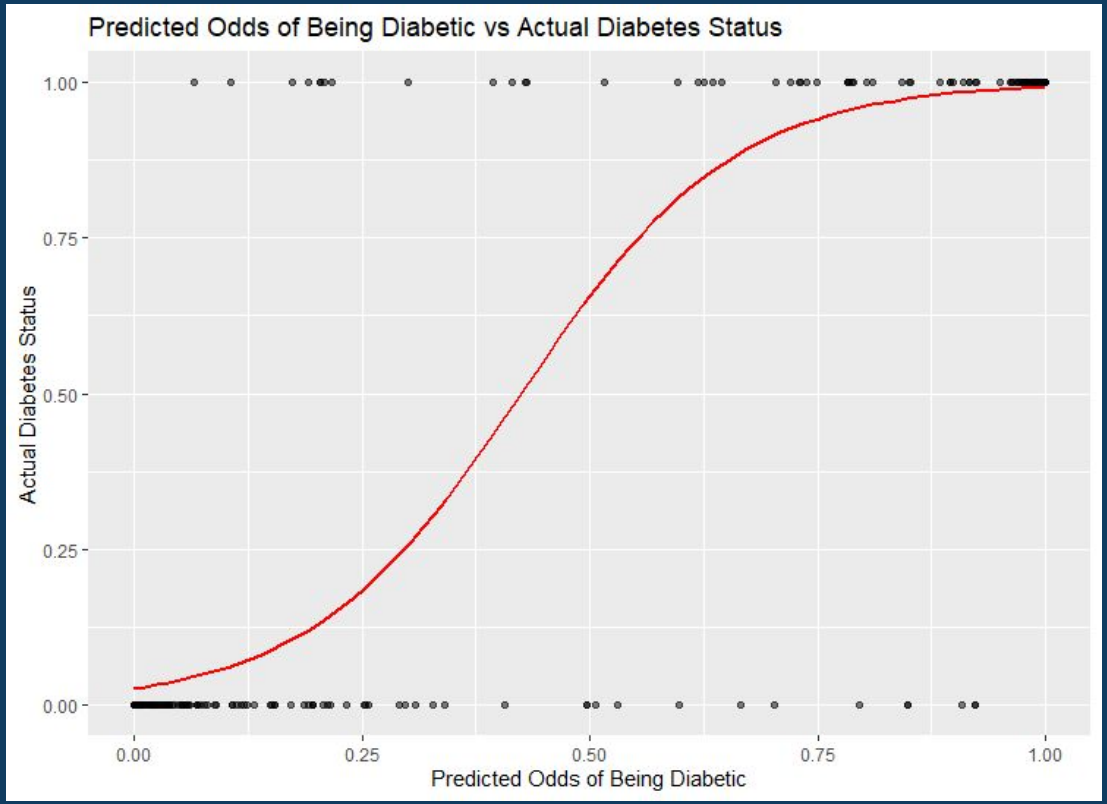
CONFUSION MATRICES

Pre-Predicted Individuals

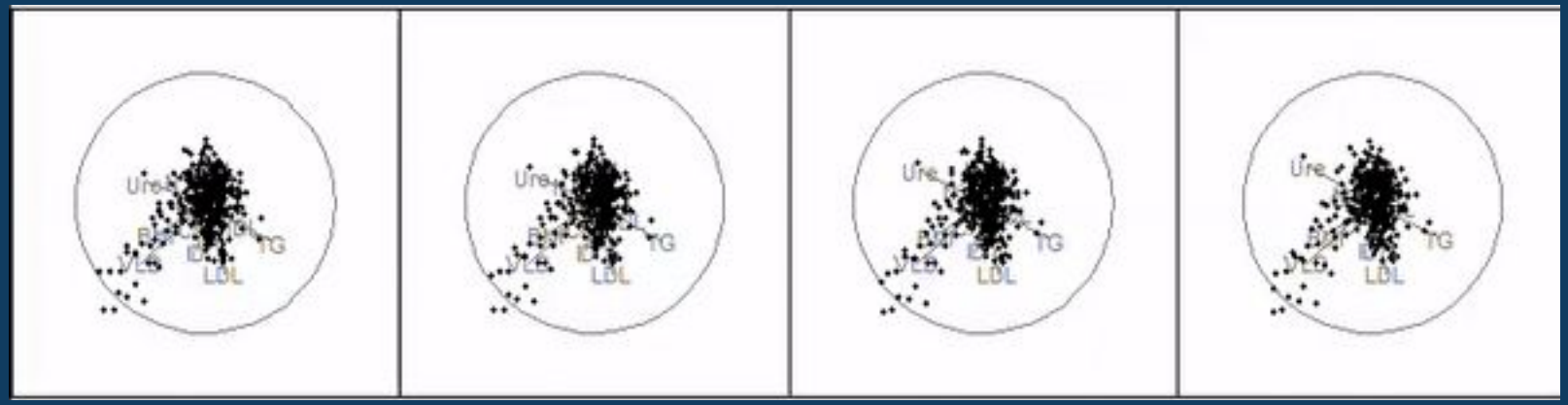
	Actually Not Diabetic	Actually Diabetic
Predicted Not Diabetic	0	13
Predicted Diabetic	0	40

Misclassification Rate of 0.245

GRAPH


Predicted Odds of Being Diabetic vs Actual Diabetes Status
Actual Diabetes Status
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Predicted Odds of Being Diabetic

K-NEAREST NEIGHBOR PLOT

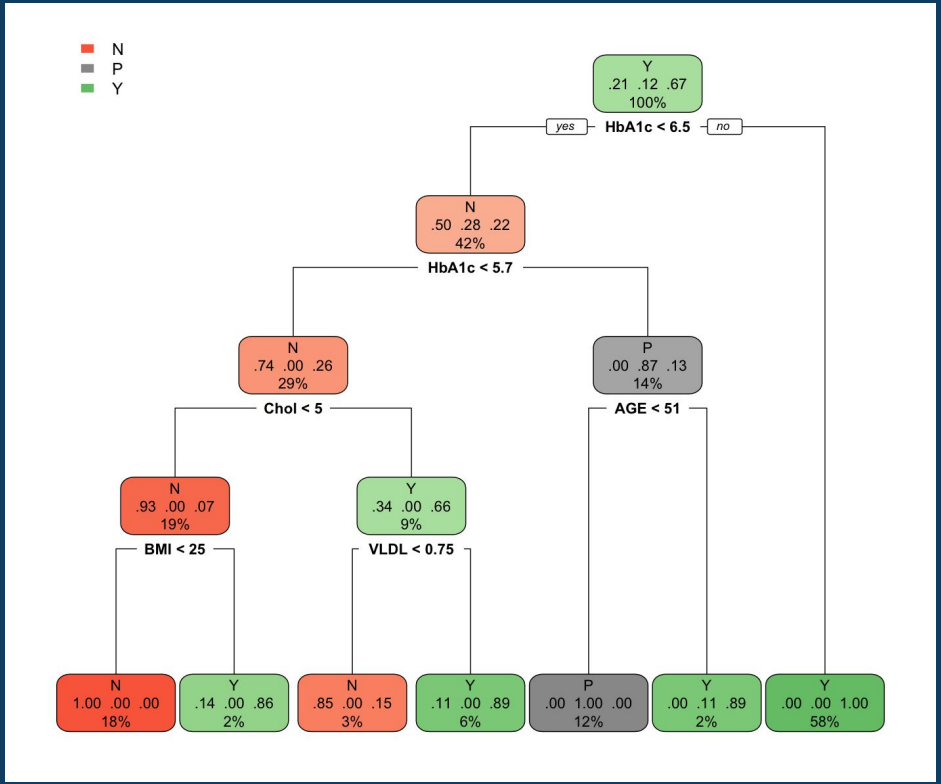


K-NEAREST NEIGHBOR PLOT

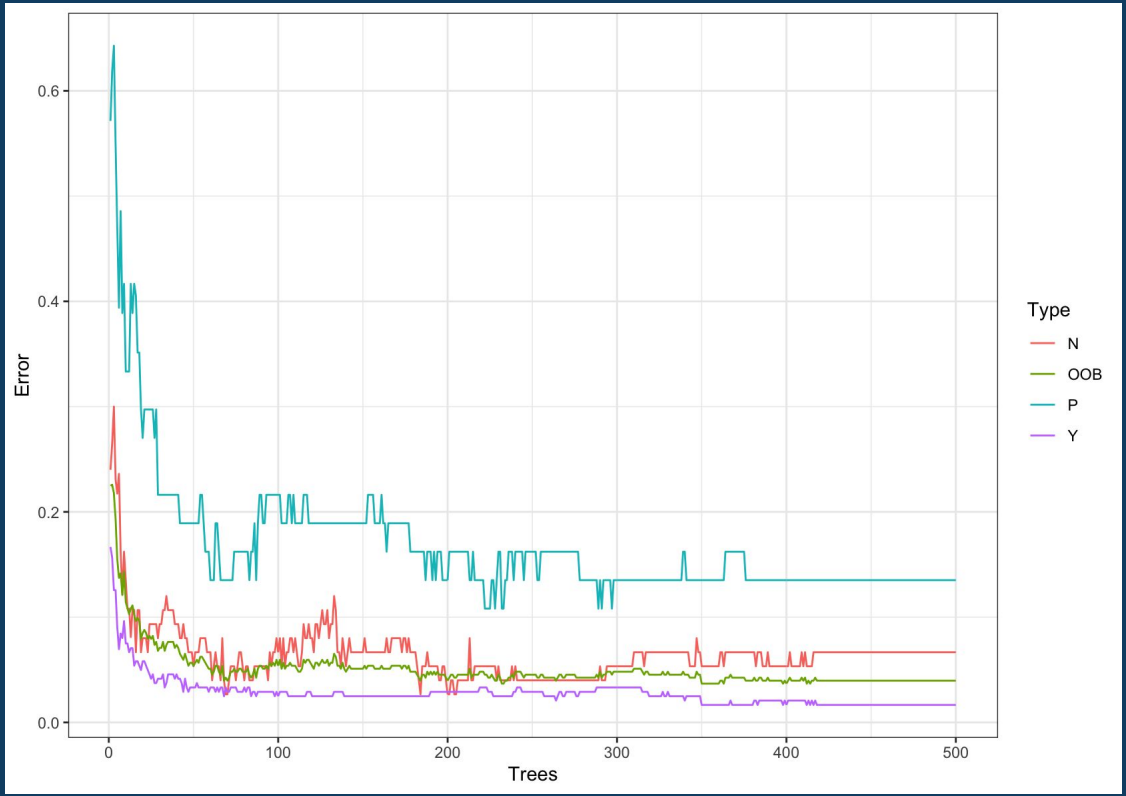
```r
##Generate a random number that is 90% of the total number of rows in dataset.
ran <- sample(1:nrow(Age.55.above), 0.9 * nrow(Age.55.above))

##the normalization function is created
nor <-function(x) { (x -min(x))/(max(x)-min(x))   }

##Run nomalization on first 4 coulumns of dataset because they are the predictors
iris_norm <- as.data.frame(lapply(Age.55.above[,-c(1,2,3,14)], nor))

##extract training set
iris_train <- iris_norm[ran,]
##extract testing set
iris_test <- iris_norm[-ran,]
##extract 5th column of train dataset because it will be used as 'cl' argument in knn function.
iris_target_category <- Age.55.above[ran,14]
##extract 5th column if test dataset to measure the accuracy
iris_test_category <- Age.55.above[-ran,14]
##load the package class
library(class)
##run knn function
pr <- knn(iris_train,iris_test,cl=iris_target_category,k=3)

##create confusion matrix
tab <- table(pr,iris_test_category)

##this function divides the correct predictions by total number of predictions that tell us how accurate teh mode
l is.
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(tab)
```

```
## [1] 84.44444
```

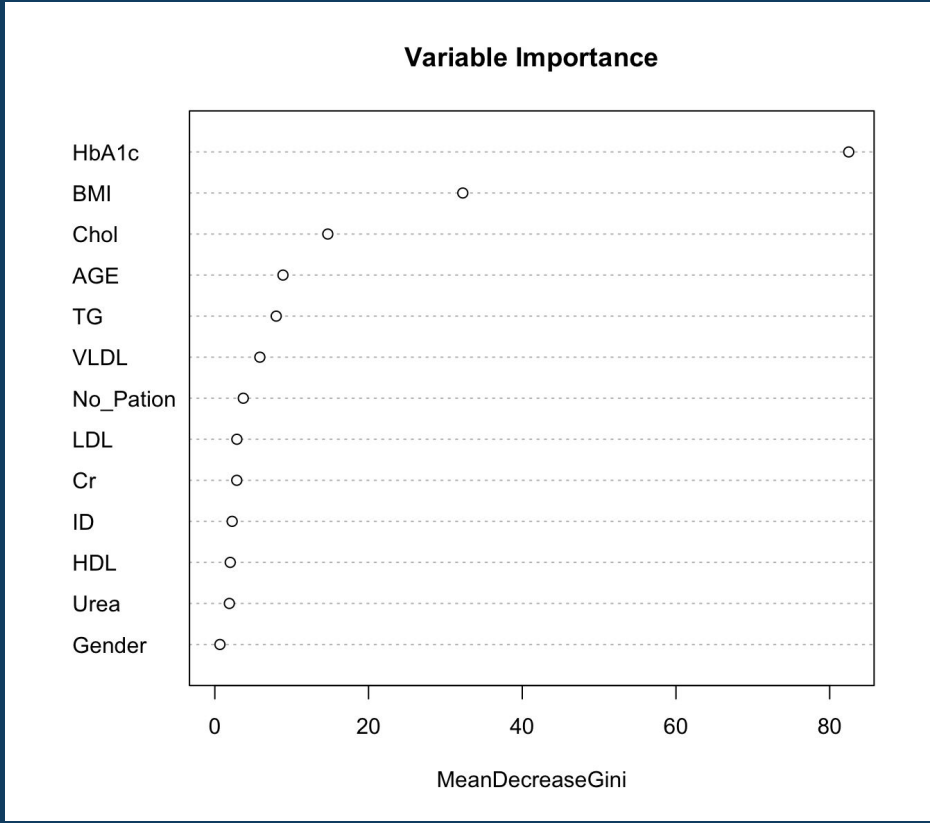
CLASSIFICATION TREE



RANDOM FOREST



RANDOM TREE


Variable Importance
HbA1c
BMI
Chol
AGE
TG
VLDL
No_Pation
LDL
Cr
ID
HDL
Urea
Gender
0
20
40
60
80
MeanDecreaseGini

EVALUATION

- We found 2 models within the scatter plot
- Determined that VLDL > 5 has a high correlation with multiple predictors
- Logistic Regression was affected heavily by the imbalance of the data but for predictive purposes was still extremely effective
- K-nearest neighbor was used to predict but didn't do a good job as expected
- Random forest was used which gave us a good error rate
- We can predict diabetes status with 0.022% error rate

THANK YOU

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/