# STAT - 515 FINAL PROJECT

## Health Prediction Modeling and Analysis

Group-5

Puneeth Velivela, Hemanth Reddy Kurapati

STAT-515-002(FALL 2023)

Prof.Dr.Isuru Dassanayake

December 7, 2023

**ABSTRACT:**

This study explores factors influencing the probability that a patient will receive a diabetes diagnosis by conducting a comprehensive analysis of a diabetes dataset. We examine the connections between different clinical variables and the 'CLASS' variable, which indicates the diagnosis status (N: Normal, P: Pre-diabetic, Y: Diabetic), by utilizing an extensive range of statistical and machine learning techniques.

Preprocessing of the data, exploratory data analysis (EDA), and modeling methods like logistic regression and decision tree modeling are all included in the analysis. We examine the effects of age, gender, BMI, and biochemical markers on the diagnosis of diabetes, attempting to identify significant characteristics in predictive models.

Furthermore, we investigate the connection between different factors and HbA1c levels using hierarchical clustering and regression models. In order to provide insight into variable correlations and predictive power, the relative merits of ridge regression, lasso regression, and linear regression are evaluated. In addition, we examine the connection between age and BMI, utilizing regression analysis and hypothesis testing to determine whether a meaningful correlation exists. The results underline the importance of specific clinical markers and offer insights into predictive modeling, thereby contributing to a more nuanced understanding of the factors influencing diabetes diagnosis.

**INTRODUCTION:**

Globally, the prevalence of diabetes, a complex metabolic disorder with multifactorial origins, is rising. Effective healthcare interventions depend on an understanding of the complex interactions between clinical variables in the diagnosis of diabetes. With the goal of revealing the complex connections between patient characteristics and diabetes diagnosis, we delve deeply into a diabetes dataset in this study.

Our research is centered on the 'CLASS' variable, which denotes the patients' diagnostic status (Normal, Pre-diabetic, and Diabetic). To analyze the dataset, we employ a wide range of statistical and machine learning methods, starting with data preprocessing to guarantee data integrity. Our study includes a thorough exploratory data analysis that clarifies the distribution and connections between clinical variables.

To determine the influence of age, gender, BMI, and different biochemical markers on the diagnosis of diabetes, logistic regression models are utilized. To find significant features, decision tree modeling is also used, offering a comparison analysis with logistic regression. We also go into the domain of HbA1c levels, which is a vital sign of glycemic control. Regression models are utilized to assess the predictability of HbA1c levels and identify correlation patterns. These models include ridge regression, lasso regression, and linear regression.

In addition, we study the association between age and BMI, looking for possible correlations using regression analysis and hypothesis testing. The goal of this thorough analysis is to provide important insights into the variables influencing diabetes diagnosis, providing a basis for well-

informed healthcare decision-making. The methods and findings are covered in detail in the following sections.

**Research Questions:**

1) a) How do various factors contribute to the likelihood of patients being diagnosed with diabetes, as indicated by the 'CLASS' variable.
   b) Identify and analyze the most influential features identified by the decision tree model. Assess the comparative effectiveness of a decision tree model against logistic regression in determining the presence or absence of a medical condition.
2) What is the relationship between hbA1c levels and various factors, and how well can hbA1c be predicted using regression models? Additionally, can hierarchical clustering provide insights into the patterns of correlation among hbA1c and other variables?
3) H0: There is no relationship between AGE and BMI
   versus the alternative hypothesis
   Ha: There is some relationship between AGE and BMI
   Test: Trying to find out the correlation between AGE and BMI and plot the relation.

**Dataset:**

The dataset selected is a diabetes dataset and is taken from Mendeley data, It has 14 columns and 1000 rows. Here, mentioned below is the first 20 rows of the dataset.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | No_Pation | Gender | AGE | Urea | Cr | HbA1c | Chol | TG | HDL | LDL | VLDL | BMI | CLASS | |
| 2 | 1 | 34325 | M | 58 | 20.8 | 800 | 9.1 | 6.6 | 2.9 | 1.1 | 4.3 | 1.3 | 33 | Y | |
| 3 | 1 | 34325 | M | 56 | 20.8 | 800 | 9 | 4.6 | 2 | 1.2 | 2.5 | 0.9 | 35 | Y | |
| 4 | 2 | 44835 | M | 60 | 2.1 | 56 | 7.6 | 3.3 | 1.7 | 0.9 | 1.7 | 0.8 | 36.6 | Y | |
| 5 | 2 | 44835 | M | 60 | 2.1 | 56 | 7.2 | 6.3 | 3.7 | 1 | 3.7 | 1.7 | 28 | Y | |
| 6 | 3 | 23972 | F | 56 | 4 | 45 | 9.2 | 4.1 | 0.6 | 1.3 | 1.4 | 0.9 | 30 | Y | |
| 7 | 3 | 41248 | M | 53 | 4.2 | 55 | 8 | 5 | 2.5 | 1.3 | 2.3 | 1.6 | 30 | Y | |
| 8 | 4 | 34301 | F | 43 | 2.1 | 55 | 5.7 | 4.7 | 5.3 | 0.9 | 1.7 | 2.4 | 25 | P | |
| 9 | 4 | 18642 | M | 55 | 9.2 | 101 | 8.5 | 5.6 | 1.9 | 1.3 | 1.2 | 0.7 | 28 | Y | |
| 10 | 5 | 35150 | M | 63 | 7 | 84 | 8.1 | 6 | 2.2 | 1.1 | 4 | 1 | 28 | Y | |
| 11 | 5 | 51477 | M | 58 | 5.9 | 67 | 9.9 | 5.2 | 1.8 | 0.9 | 3.6 | 0.6 | 29 | Y | |
| 12 | 6 | 23973 | F | 61 | 5.1 | 72 | 11.5 | 4.4 | 2.1 | 1.1 | 2.5 | 0.9 | 26 | Y | |
| 13 | 6 | 45308 | F | 58 | 6 | 66 | 6.9 | 5.7 | 1.3 | 1.4 | 4.9 | 0.6 | 24 | Y | |
| 14 | 7 | 21354 | M | 44 | 6.8 | 64 | 4.9 | 4.9 | 2.8 | 2 | 1.8 | 1.2 | 21 | N | |
| 15 | 7 | 34278 | F | 46 | 3 | 59 | 5.1 | 5.7 | 3.8 | 1.3 | 2.8 | 1.7 | 24 | Y | |
| 16 | 8 | 45703 | M | 51 | 3.9 | 53 | 10.9 | 3.6 | 1.1 | 0.8 | 2.3 | 1 | 29 | Y | |
| 17 | 8 | 79133 | M | 55 | 5.8 | 60 | 9 | 4.6 | 1.9 | 1.2 | 2.6 | 0.8 | 30 | Y | |
| 18 | 9 | 85922 | F | 35 | 3.9 | 38 | 5.4 | 3.8 | 5.9 | 0.5 | 4.3 | 1 | 22 | N | |
| 19 | 9 | 23974 | F | 60 | 6 | 72 | 10.7 | 4.4 | 2.1 | 1.1 | 2.5 | 0.9 | 26 | Y | |
| 20 | 10 | 568412 | M | 40 | 5 | 63 | 4 | 4.8 | 2.5 | 1.1 | 2.7 | 1.1 | 23 | N | |

Dataset

*Figure 1: Dataset*

**Data Preprocessing:**

Data preprocessing is a crucial step in the data analysis pipeline that involves cleaning and transforming raw data into a format suitable for analysis. Below are the cleaning steps we have done to the dataset.

**Handling Missing Values:**

We used the na.omit() function to remove rows with missing values. This ensures that the dataset is free from instances where essential information is incomplete.

**Feature Selection:**

We removed unnecessary columns, such as patient IDs that do not contribute to the analysis or contain redundant information. This helps in reducing the dimensionality of the dataset and focusing on relevant features.

**Data Type Conversion:**

Converted the 'Gender' variable to a numeric format for standardization.

Ensured that the 'CLASS' variable is a factor with three levels ('N', 'P', 'Y'), which is essential for classification tasks.

**Handling Categorical Data:**

Normalized gender values between 0 and 1, representing 'F' and 'M', respectively. This ensures uniformity and compatibility with numerical variables.

**Removing Extra Spaces:**

Trimmed extra spaces in the 'CLASS' variable, ensuring consistency and eliminating potential issues related to leading or trailing spaces.

**Standardization of Numerical Variables:**

Standardized or normalized numerical variables like 'AGE', 'Urea', 'Cr', 'HbA1c', 'Chol', 'TG', 'HDL', 'LDL', 'VLDL', and 'BMI'. Standardization is crucial when using certain machine learning algorithms that are sensitive to the scale of input features.

In summary, our data preprocessing steps focus on ensuring data quality, handling missing values, optimizing feature selection, and preparing the data for subsequent analysis. These steps contribute to a cleaner and more reliable dataset, setting the stage for meaningful exploration and modeling.

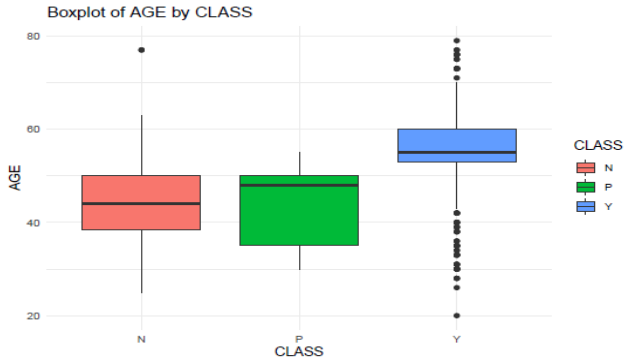
**Visualizations:**

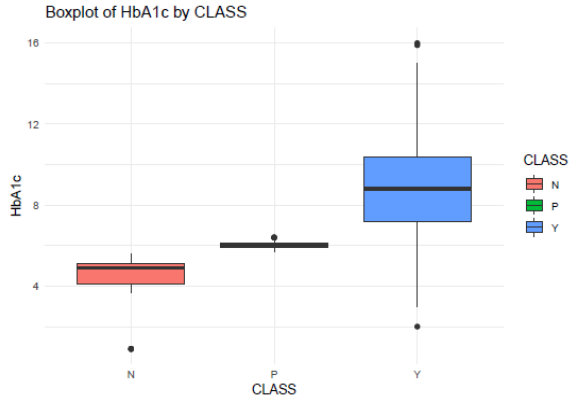
Visualization is a basic technique for producing an understandable visual depiction of an upcoming event. We can picture the event and practice for it in advance to make sure we are sufficiently ready. You can develop the confidence you need to succeed by visualizing your successes. The four fundamental elements of visualization are relationship, composition, comparison, and distribution. To make our data easier to understand, use visual aids such as histograms, maps, charts, graphs, and more. We can quickly identify patterns, trends, and outliers in our data-thanks to-data visualization. To ensure that the data has been entered into the system correctly, the first

few rows of the dataset can be seen after it has been loaded into our application tool. Below are the few visualizations what we have done to the dataset.

**Boxplots by Class:**

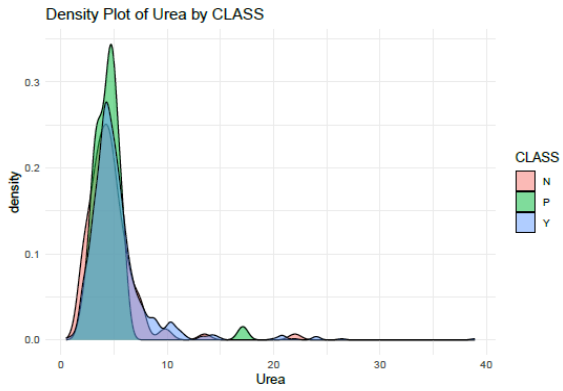
We created boxplots for each numerical variable in the dataset, grouping the data by the 'CLASS' variable. These boxplots provide a visual summary of the distribution of each variable across different classes ('N', 'P', 'Y'). It helps identify potential patterns or differences in the data distribution among the different classes.

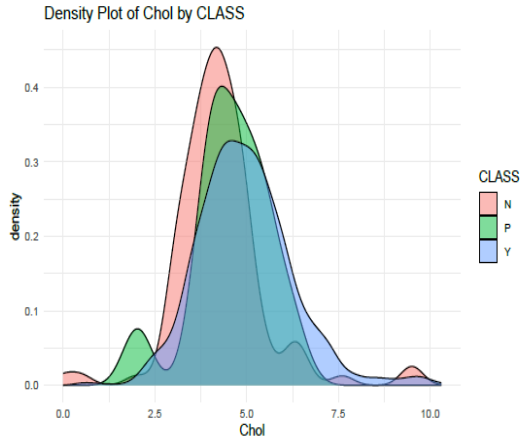


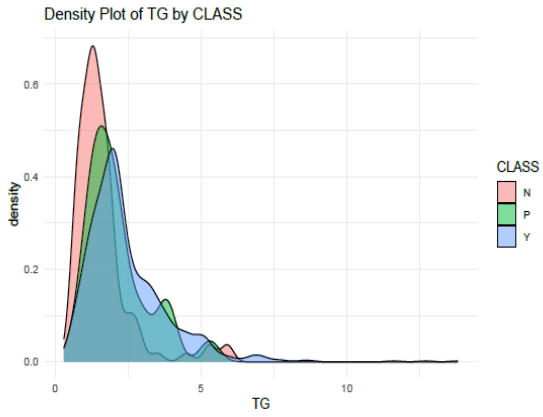




**Density Plots by Class:**

Density plots were generated for each numerical variable, with colors representing different classes. These plots illustrate the distribution of values within each class, offering insights into the shape and overlap of the distributions. They are particularly useful for understanding the underlying probability distribution of the data.

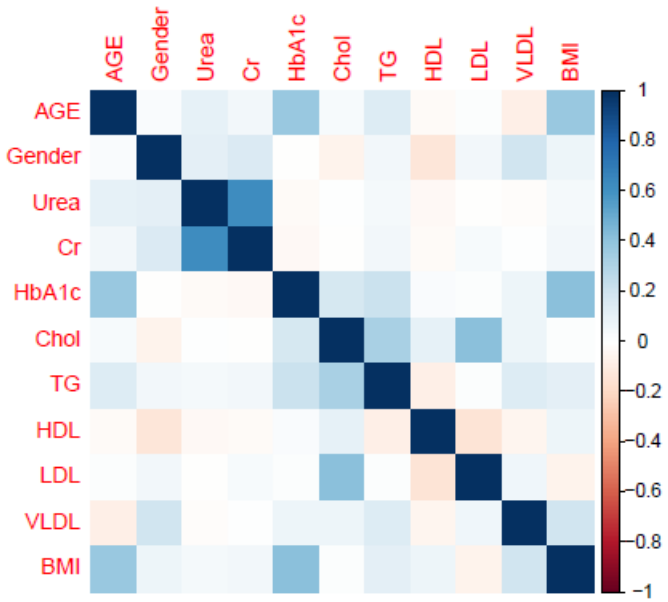



## Correlation Matrix and Hierarchical Clustering:

A correlation matrix was created to explore relationships between numerical variables. The corrplot visually represents the strength and direction of these correlations. Hierarchical clustering using complete linkage was applied to reveal patterns of similarity among observations.




**Research Questions:**

**Question1: a) How do various factors contribute to the likelihood of patients being diagnosed with diabetes, as indicated by the 'CLASS' variable?**

**b) Identify and analyze the most influential features identified by the decision tree model. Assess the comparative effectiveness of a decision tree model against logistic regression in determining the presence or absence of a medical condition.**

We used logistic regression statistical modeling to evaluate the variables that influence a patient's chance of receiving a diabetes diagnosis. The association between the response variable 'CLASS' (which indicates the presence or absence of diabetes) and a group of predictor variables, such as AGE, BMI, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, and VLDL, was investigated using the logistic regression model.

The direction and strength of the relationships between these factors and the likelihood of receiving a diabetes diagnosis are revealed by the coefficients derived from the logistic regression model. Positive coefficients, on the other hand, suggest a positive association and negative coefficients a negative association with the risk of diabetes.

```
## Call:
## multinom(formula = CLASS ~ AGE + Gender + BMI + HbA1c + Chol +
##     TG + LDL + VLDL, data = train_data)
##
## Coefficients:
##   (Intercept)         AGE    Gender       BMI    HbA1c      Chol        TG
## P   -21.16174 -0.02571233 1.1801140 0.4474288 1.608831 0.4690297 0.6718180
## Y   -43.39863  0.04701047 0.7297656 0.9034049 2.246052 1.0751776 0.7806315
##          LDL       VLDL
## P -0.3593638 -0.1657575
## Y  0.1958714  0.1124938
##
## Std. Errors:
##   (Intercept)        AGE    Gender       BMI     HbA1c      Chol        TG
## P    4.560205 0.02925444 0.5171667 0.1426622 0.3190284 0.2704868 0.2677504
## Y    5.265610 0.03019676 0.5589079 0.1528250 0.3334375 0.2819974 0.2774924
##         LDL      VLDL
## P 0.3067277 0.3166076
## Y 0.3136195 0.3102269
```



```
##
## Residual Deviance: 261.5424
## AIC: 297.5424
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   N   P   Y
##          N  16   2   8
##          P   0   4   3
##          Y   0   3 164
##
## Overall Statistics
##
##               Accuracy : 0.92
##                 95% CI : (0.8733, 0.9536)
##    No Information Rate : 0.875
##    P-Value [Acc > NIR] : 0.02920
##
##                  Kappa : 0.6892
##
##  Mcnemar's Test P-Value : 0.01857
##
```

The residual Deviance from the logistic regression model is 261.54 and got an accuracy of 92% when considered the full by considering all the models.

Then we used step wise selection to find the best subset of variables and below are the results:

```
## Call:
## multinom(formula = CLASS ~ AGE + Gender + BMI + Urea + Cr + HbA1c +
##     Chol + TG + HDL + LDL + VLDL, data = train_data)
##
## Coefficients:
##   (Intercept)         AGE   Gender       BMI         Urea          Cr    HbA1c
## P   -21.11351 -0.02605882 1.266894 0.4521690 -0.122871294 0.003211243 1.635962
## Y   -44.22180  0.04699452 0.732504 0.9131893  0.000953907 0.002698984 2.253508
##        Chol        TG        HDL        LDL       VLDL
## P 0.4552685 0.6859771 0.08709528 -0.3660598 -0.1829501
## Y 1.0870918 0.7932568 0.24764479  0.1965272  0.1233282
##
## Std. Errors:
##   (Intercept)        AGE    Gender       BMI      Urea          Cr     HbA1c
## P    4.766604 0.02961927 0.5527453 0.1451605 0.1428659 0.009847796 0.3286852
## Y    5.523753 0.03048132 0.5896579 0.1556562 0.1592898 0.010993382 0.3400208
##        Chol        TG       HDL       LDL      VLDL
## P 0.2759726 0.2719025 0.5318567 0.3205775 0.3168279
## Y 0.2897455 0.2899790 0.6671583 0.3281228 0.3096093
##
## Residual Deviance: 259.4241
## AIC: 307.4241
```

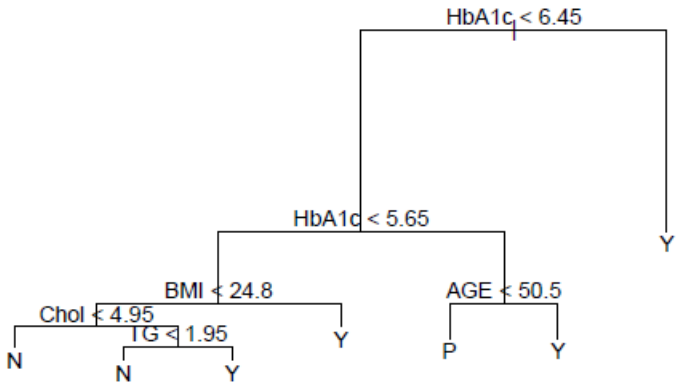
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   N   P   Y
##          N  16   2   8
##          P   0   4   2
##          Y   0   3 165
##
## Overall Statistics
##
##                Accuracy : 0.925
##                  95% CI : (0.8793, 0.9574)
##     No Information Rate : 0.875
##     P-Value [Acc > NIR] : 0.01637
##
##                   Kappa : 0.7038
##
##  Mcnemar's Test P-Value : 0.01694
##
## Statistics by Class:
```

When compared between the both the models model with best subset of variables gave the better accuracy and less Residual deviance values. So in logistic regression models the best model is with the subset selection model.

**Part B :**

We used a decision tree model to find and examine the most significant features. When it comes to capturing intricate interactions and nonlinear relationships in data, decision trees are especially useful. We can determine the most important characteristics in predicting the presence or absence of diabetes by looking at the decision tree's structure and the variables chosen at each split.

In some cases, the decision tree model is superior to logistic regression in terms of comparison. Logistic regression might have trouble capturing interactions and non-linearity's; decision trees can. Additionally, they offer a visual depiction of decision-making procedures, which facilitates the interpretation and dissemination of outcomes.

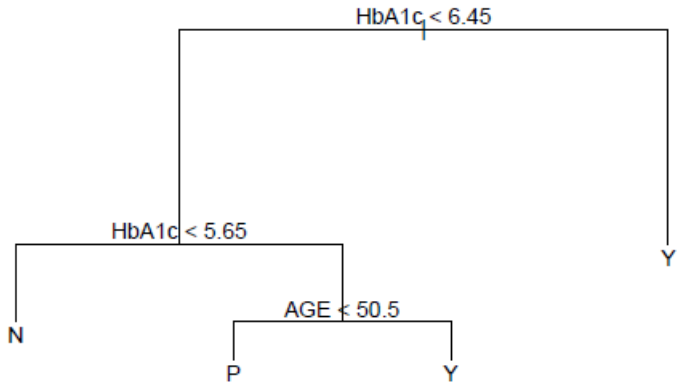



From the above figure we can interpret that HbA1c is the important variable based on this the value of diabetes class will vary if it is less than 6.45 then the tree follow left path and if it is greater than it follows right part. The second level of the tree is also HbA1c values and it follows different paths based on its values. The third important variable in the dataset is the BMI and AGE and then Chol and TG values staying at the last nodes of the tree. The accuracy of the predictions generated from this tree is 98.5%

```
predictions_unpruned <- predict(unpruned_tree, newdata = test_data, type = "class")

confusionMatrix(predictions_unpruned,test_data$CLASS)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   N   P   Y
##          N  16   0   3
##          P   0   9   0
##          Y   0   0 172
##
## Overall Statistics
##
##                Accuracy : 0.985
##                  95% CI : (0.9568, 0.9969)
##     No Information Rate : 0.875
##     P-Value [Acc > NIR] : 1.075e-08
##
##                   Kappa : 0.9369
```

Then let's prune the tree by cutting of some branches and check if accuracy increases.




```
predictions_pruned <- predict(pruned_tree, newdata = test_data, type = "class")

confusionMatrix(predictions_pruned,test_data$CLASS)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   N   P   Y
##          N  16   0  12
##          P   0   9   0
##          Y   0   0 163
##
```



```
## Overall Statistics
##
##                Accuracy : 0.94
##                  95% CI : (0.8975, 0.9686)
##     No Information Rate : 0.875
##     P-Value [Acc > NIR] : 0.001914
##
##                   Kappa : 0.7807
##
##  Mcnemar's Test P-Value : NA
##
```

The accuracy didn't improve with the full tree so for this dataset the unpruned tree gives us the better result.

Now in answering to the research question the tree model gave us better results when compared with the multinomial logistic regression.

**Question 2) What is the relationship between hbA1c levels and various factors, and how well can hbA1c be predicted using regression models? Additionally, can hierarchical clustering provide insights into the patterns of correlation among hbA1c and other variables.**

Here, we are performing a comprehensive analysis of the dataset to understand the relationship between hbA1c levels and various factors. hbA1c is the important factor in predicting diabetes, so we are trying to find the relation between hbA1c levels with other factors.

We are using Regression models to find out which model is better to predict the hbA1c value. We are performing linear regression, ridge and lasso regression and finding out which is giving better performance.

```
> # Regression Analysis
> lm_model <- lm(HbA1c ~ ., data = data)
> lm_predictions <- predict(lm_model, newdata = data)
> print(lm_mse)
[1] 4.065891
```

From linear regression, we got the mse value as “4.06”.

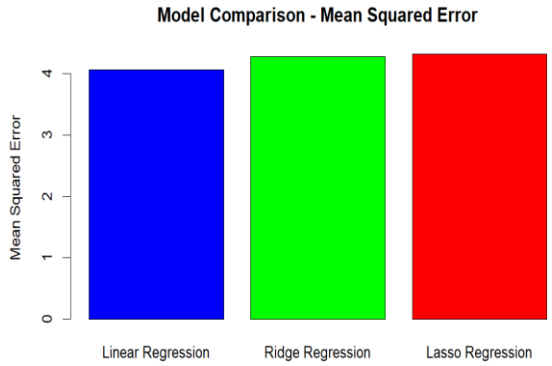
```
> # Ridge Regression
> ridge_model <- cv.glmnet(as.matrix(data[, names(data) != "HbA1c"]), data$HbA1c, alpha = 0)
> ridge_predictions <- predict(ridge_model, newx = as.matrix(data[, names(data) != "HbA1c"]))
> ridge_mse <- mean((ridge_predictions - data$HbA1c)^2)
> print(ridge_mse)
[1] 4.281578
```

From Ridge regression, we got the mse value as “4.28”.

```
> # Lasso Regression
> lasso_model <- cv.glmnet(as.matrix(data[, names(data) != "HbA1c"]), data$HbA1c, alpha = 1)
> lasso_predictions <- predict(lasso_model, newx = as.matrix(data[, names(data) != "HbA1c"]))
> lasso_mse <- mean((lasso_predictions - data$HbA1c)^2)
> print(lasso_mse)
[1] 4.314229
```

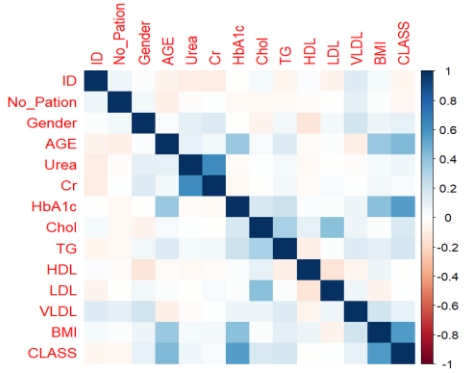
From Lasso regression, we got the mse value as “4.31”.

After getting the mse values for all three regressions, we are plotting a bar graph to observe which regression is best.



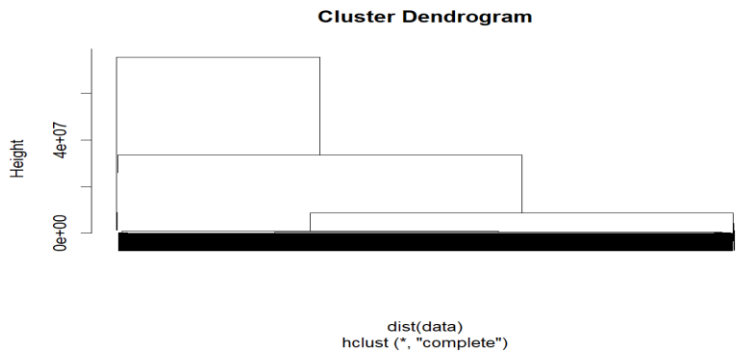

From the bar plot, we can say that Linear regression fits better, as the value for this is less compared to other two regression models.

Now, we are plotting a correlation matrix where we can find what factors are directly dependent.




We can see that HDL and Gender , HDL and LDL having high correlation. We did this using corrplot package.

Now, we are finding whether hierarchical clustering provide insights into patterns with correlations among hbA1c and other variables.



But, the cluster dendogram doesnot provide a good visualization as variables are overlapping at the bottom.

**Question 3) H0: There is no relationship between AGE and BMI**

**versus the alternative hypothesis**

**Ha: There is some relationship between AGE and BMI**

**Test: Trying to find out the correlation between AGE and BMI and plot the relation.**

Here, we using hypothesis tests to determine whether there is relationship between AGE and BMI. H0 is null hypothesis which says there is no relationship between them. Ha is alternative hypothesis which says there is relationship between them.

First, we are calculating the correlation coefficient between "AGE" and "BMI", this provides a quantitative measure for linear relationship between variables.

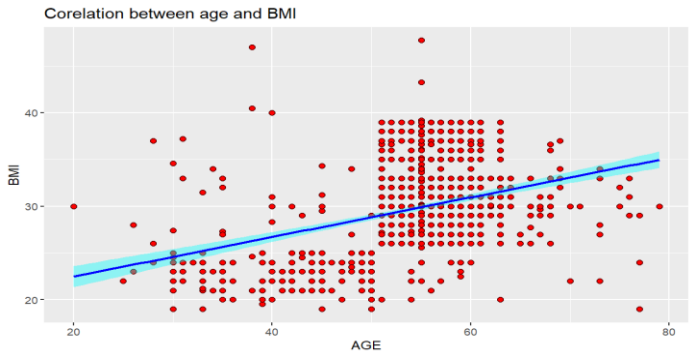
```
> cor(data$AGE,data$BMI)
[1] 0.375956
```

The correlation coefficient obtained between them is 0.37.

Next, we are finding out the summary of the fit we made.

```
> summary(fit)

Call:
lm(formula = AGE ~ BMI, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.809  -3.276   0.524   4.191  30.524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.81010    1.55988   21.68   <2e-16 ***
BMI          0.66664    0.05201   12.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.158 on 998 degrees of freedom
Multiple R-squared:  0.1413,    Adjusted R-squared:  0.1405
F-statistic: 164.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

At last, we are plotting the test. We are fitting to 99%.




From this, we can say that our model suggests a statistically significant relationship between "AGE" and "BMI".

**Future Research:**

1) **Advanced Predictive Modeling:**

More sophisticated predictive modeling approaches, like ensemble methods or deep learning, may be investigated in future studies to improve the accuracy of diabetes diagnosis predictions. Ensemble models, such as gradient boosting or random forests, have the potential to capture intricate relationships and nonlinearities that conventional models might miss.

2) **Feature Importance Analysis:**

Learn more about feature importance analysis, particularly as it relates to decision tree modeling. Healthcare practitioners can focus on important factors during patient assessments by having a clear understanding of the hierarchy of influential features in diabetes prediction.

3) **Integration of Biomarkers:**

Examine how to incorporate genetic information or other biomarkers into the analysis. This could lead to the discovery of new risk factors or indicators that influence the diagnosis of diabetes, opening the door to more specialized and focused interventions.

**Conclusion:**

To sum up, this study offers insightful information about the variables affecting the diagnosis of diabetes. The successful identification of significant features for diabetes status prediction was achieved through the combination of decision tree modeling and logistic regression. The research highlights the significance of particular clinical indicators, such as age, BMI, and HbA1c, in comprehending and forecasting diabetes.

Regression models were used to analyze HbA1c levels, demonstrating the predictive power of linear regression and providing a benchmark for comparisons with more complex models in the future. A more complex understanding of the relationships between variables is facilitated by the investigation of correlation patterns using hierarchical clustering.

As we advance, we will be able to more accurately predict, diagnose, and intervene in diabetes by utilizing sophisticated modeling techniques, incorporating longitudinal data, and investigating new

biomarkers. Future efforts to enhance diagnostic precision, improve predictive models, and eventually advance personalized healthcare for people with or at risk from diabetes will build on the research presented here.

**References:**

[1] Dataset: Rashid, A. (2020). Diabetes Dataset. Data.mendeley.com, [Online]: Available https://doi.org/10.17632/wj9rwkp9c2.

[2] Kshitiz Sirohi. (2018, December 24). Simply Explained Logistic Regression with Example in R. Medium; Towards Data Science. https://towardsdatascience.com/simply-explained-logistic-regression-with-example-in-r-b919acb1d6b3

[3] Multinomial Logistic Regression with R | 1. Data and Model. (n.d.). Www.youtube.com. Retrieved December 7, 2023, from https://www.youtube.com/watch?v=S2rZp4L_nXo

[4] Starkweather, J., & Moske, A. (n.d.). Multinomial Logistic Regression. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/Benchmarks/MLR_JDS_Aug2011.pdf

[5] Diabetes: Diabetes Dataset in heplots: Visualizing Hypothesis Tests in Multivariate Linear Models. (n.d.). Rdrr.io. Retrieved December 7, 2023, from https://rdrr.io/rforge/heplots/man/Diabetes.html