Ashwati Nair
HI 2021-1300
8/9/2023

Does Elevated Blood Glucose Increase the Risk of Heart Disease?

Heart disease remains one of the leading causes of death. There are many risk factors for heart disease, which can be caused by lifestyle (smoking, lack of exercise), genetic predisposition, and comorbidities. One major comorbidity is diabetes mellitus, which is categorized by elevated blood glucose levels. Elevated blood glucose levels, as well as insulin resistance can cause damage to the cardiovascular system in many ways. Thus, it is important to establish the relationship between glucose and cardiovascular disease so that both clinicians and patients can take the appropriate measures to prevent life-threatening events such as myocardial infarction. Data analysis is an effective tool for establishing relationships between health risk factors and disease. Here, RStudio will be used to perform descriptive and inferential statistics. It will also be utilized to build machine learning models that can predict if an individual will develop heart disease based on if they have elevated blood glucose.

Many studies have been done to establish the correlation between blood glucose and heart disease. Glucose can cause harm to the blood vessels of the heart. Insulin resistance, which is correlated with increased blood glucose levels, can lead to oxidative stress, and worsen myocardial injury (Poznyak et al 2022). This applies to people with both Type 1 and Type 2 diabetes, and middle-aged adults are at the highest risk. A study done in 2019 found that "Middle-age individuals with diabetes have high long-term absolute risk for CVD" (Bancks et al). The same study also concluded that due to the risk, middle aged adults with diabetes should have their cardiovascular health monitored along with their glucose. Even before an individual is formally diagnosed with diabetes, they are at risk for heart disease. According to a 2021 study,

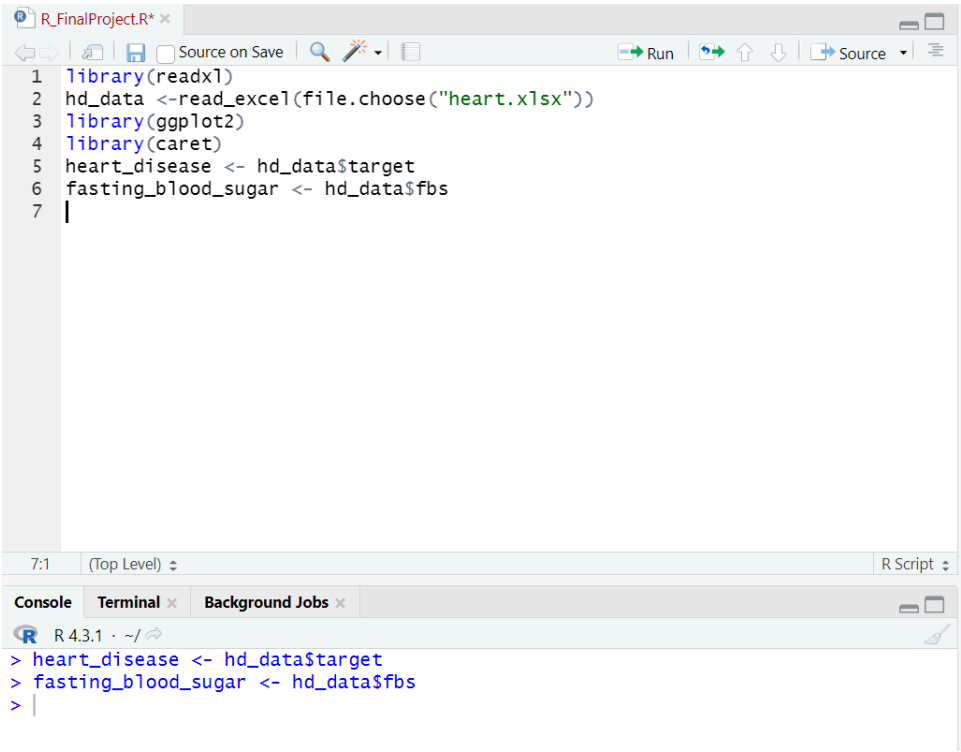
“glucose levels even below the diagnostic diabetes threshold are associated with coronary heart disease, cardiovascular mortality and total mortality. The present study shows that there also is a graded association between glucose and atrial fibrillation and heart failure, starting already at prediabetes levels” (Lind et al). This further highlights the importance of monitoring and early intervention. Diabetes was also found to decrease the effectiveness of statins, which are medications that reduce the occurrence of plaque in a person’s arteries (Mashayekhi-Sardoo et al, 2021). Increased plaque, also known as dyslipidemia, can lead to myocardial infarction. Insulin resistance can also cause changes in the formation of lipoproteins, leading to plaque formation (Eckel et al, 2022). As discussed, there is extensive literature to support the correlation between glucose and heart disease. Now, it is time to demonstrate how data analysis can determine the relationship between the two in real time.

The project's goals are to determine if there is a correlation between elevated fasting blood glucose and heart disease and to predict if a patient is at risk of developing heart disease based on their fasting glucose. The null hypothesis is that there is no correlation between elevated fasting blood glucose levels and heart disease, and the alternate hypothesis is that there is a correlation between the two variables. This will be done using the Pearson’s Chi test and finding the p value. To predict whether a patient will develop heart disease from their fasting glucose, logistic regression and random forest will be used as the algorithms and compared, with fasting blood sugar and “target” being the features. The models will be evaluated using a confusion matrix. For the project, the Heart Disease Dataset will be used for analysis. This dataset contains data from 1988. While it contains many patient attributes that are relevant to monitoring heart disease, such as angina, cholesterol, etc., fasting blood sugar and the “target” category, which states whether the patient has heart disease or not, will be the focus. A fasting

blood sugar above 120 is considered elevated and is labeled using 0 for not elevated, and 1 for elevated. For the “target” category, 0 indicates no disease, while 1 indicates presence of disease.

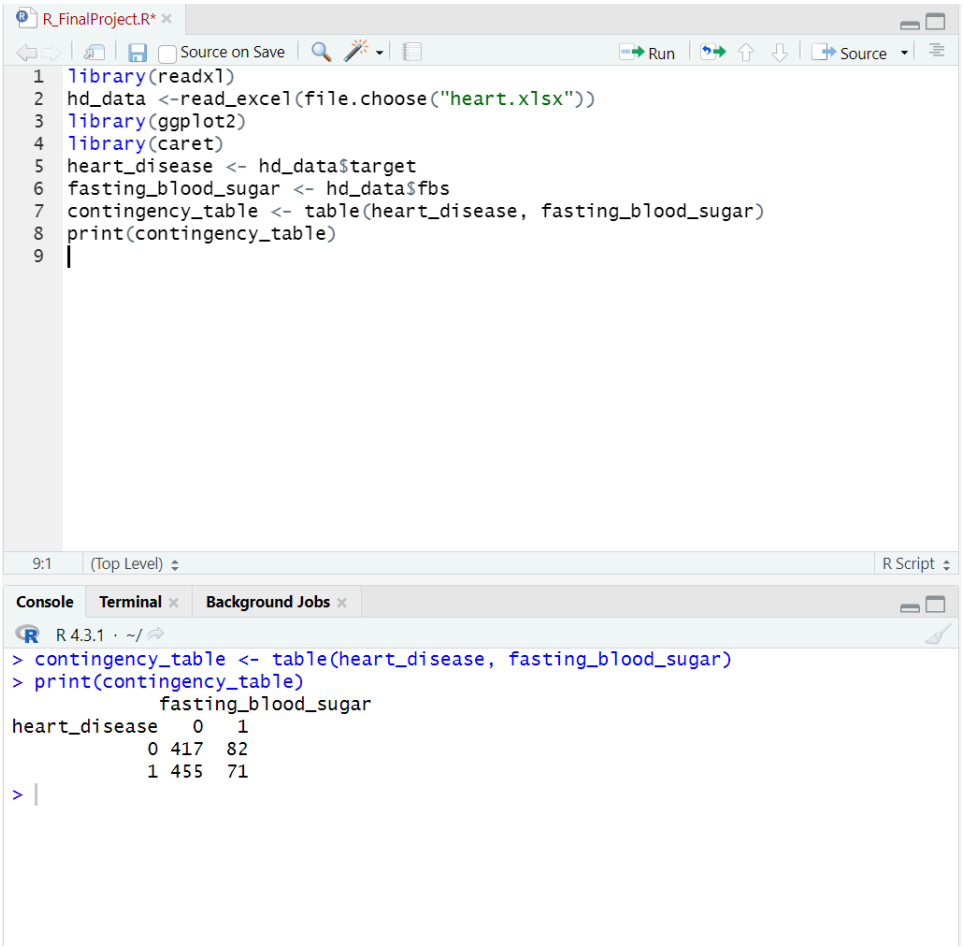
Cleaning the Data

After loading the “readxl” library into RStudio, the data from the Heart Disease dataset is loaded into the application. Since only the target and fasting blood sugar variables are needed for analysis, they were assigned individual vectors.



Descriptive Statistics and Stacked Bar Plot

Because the data for the two variables are categorical, a contingency table was created to visualize the total number of samples that fell into each category:

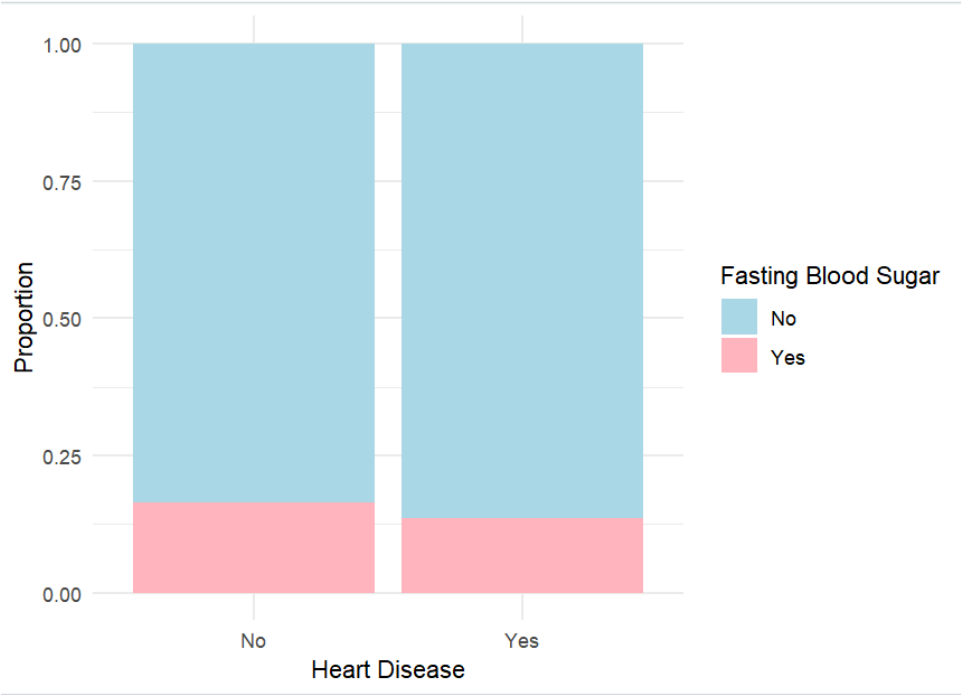


Per the table, the largest number of samples had heart disease, but did not have elevated blood sugar. The lowest number of samples had both elevated blood sugar and heart disease. To better visualize the data in the contingency table, a stacked box plot was created using ggplot2. The code for the plot is as follows:

```
data <- data.frame(heart_disease = factor(heart_disease, levels = c(0, 1), labels = c("No", "Yes")),
                   fasting_blood_sugar = factor(fasting_blood_sugar, levels = c(0, 1), labels = c("No", "Yes")))

Scattered_box_plot <- ggplot(data, aes(x = heart_disease, fill = fasting_blood_sugar)) +
```

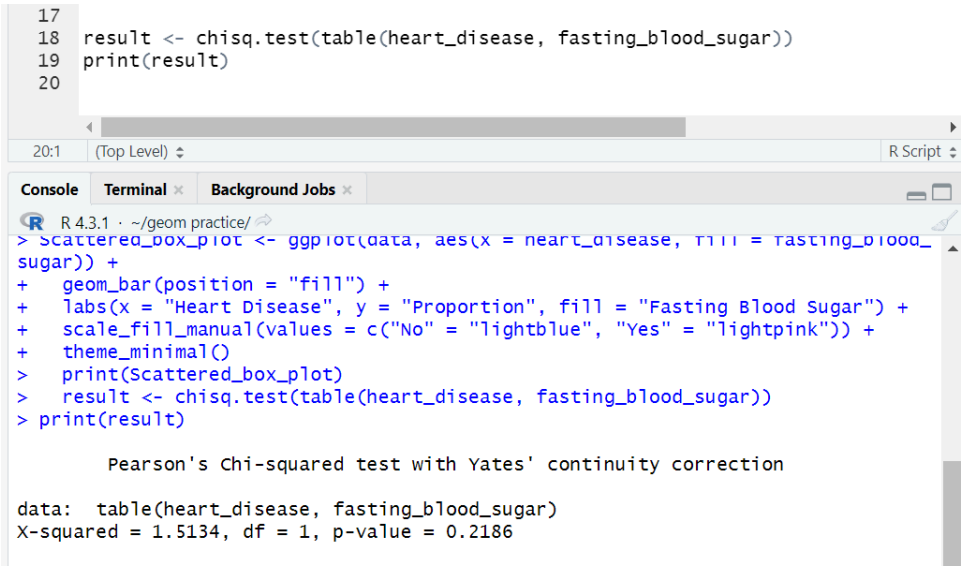
```
geom_bar(position = "fill") +
labs(x = "Heart Disease", y = "Proportion", fill = "Fasting Blood Sugar") +
scale_fill_manual(values = c("No" = "lightblue", "Yes" = "lightpink")) +
theme_minimal()
print(Scattered_box_plot)
```



Here the proportions of breakdown of the data are better seen. The most essential information to note is that less than 25% of the samples with heart disease had elevated fasting blood sugar. In fact, there were more samples who did not have heart disease with elevated fasting blood sugar. For both categories of heart disease, over 75% of samples did not have elevated fasting blood sugar.

Inferential Statistics- Pearson's Chi-squared test

For the inferential statistics portion, chi-square test is used to determine whether there is a significant association between fasting blood glucose and heart disease, as they are both categorical variables:



With a p-value of 0.2186, the result is not statistically significant at conventional significance levels (e.g., $\alpha = 0.05$). This means there is not enough evidence to reject the null hypothesis, which means there is no association between elevated fasting blood sugar and heart disease.

Machine Learning- Logistic Regression and Random Forest

Based on the data being categorical, a logistic regression model and a random forest model were created and tested so that both predictions could be compared. For both models the data was split 80/20 for testing and predictions. First the logistic regression model was examined:

```
  1 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 739, 739, 738, 739, 739, 739, ...
Resampling results across tuning parameters:

  cost  loss       epsilon  Accuracy   Kappa
  0.5   L1         0.001    0.5322656  0.04710264
  0.5   L1         0.010    0.5322656  0.04710264
  0.5   L1         0.100    0.5322656  0.04710264
  0.5   L2_dual    0.001    0.5322656  0.04710264
  0.5   L2_dual    0.010    0.5322656  0.04710264
  0.5   L2_dual    0.100    0.5322656  0.04710264
  0.5   L2_primal  0.001    0.5322656  0.04710264
  0.5   L2_primal  0.010    0.5322656  0.04710264
  0.5   L2_primal  0.100    0.5322656  0.04710264
  1.0   L1         0.001    0.5322656  0.04710264
  1.0   L1         0.010    0.5322656  0.04710264
  1.0   L1         0.100    0.5322656  0.04710264
  1.0   L2_dual    0.001    0.5322656  0.04710264
  1.0   L2_dual    0.010    0.5322656  0.04710264
  1.0   L2_dual    0.100    0.5322656  0.04710264
  1.0   L2_primal  0.001    0.5322656  0.04710264
  1.0   L2_primal  0.010    0.5322656  0.04710264
  1.0   L2_primal  0.100    0.5322656  0.04710264
  2.0   L1         0.001    0.5322656  0.04710264
  2.0   L1         0.010    0.5322656  0.04710264
  2.0   L1         0.100    0.5322656  0.04710264
  2.0   L2_dual    0.001    0.5322656  0.04710264
  2.0   L2_dual    0.010    0.5322656  0.04710264
  2.0   L2_dual    0.100    0.5322656  0.04710264
  2.0   L2_primal  0.001    0.5322656  0.04710264
  2.0   L2_primal  0.010    0.5322656  0.04710264
  2.0   L2_primal  0.100    0.5322656  0.04710264

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were cost = 0.5, loss = L1 and epsilon
 = 0.001.
```

```
31 test_data <- data[-train_indices, ]
32 print(train_data)
33 print(test_data)
34
35 fitControl <- trainControl(method = "cv", number = 10, savePredictions = TRU
36
37 lgr_model <- train(heart_disease ~ fasting_blood_sugar, data = train_data, m
38 print(lgr_model)
39
40 lgr_predictions <- predict(lgr_model, newdata = test_data)
41 print(lgr_predictions)
42
```

42:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 · ~/geom practice/

```
> lgr_predictions <- predict(lgr_model, newdata = test_data)
> print(lgr_predictions)
  [1] Yes Yes Yes No  Yes Yes No  Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes
 [19] No  Yes Yes Yes No  Yes Yes Yes Yes No  No  Yes Yes Yes Yes Yes No  Yes
 [37] No  Yes No  Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
 [55] Yes No  Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
 [73] Yes Yes Yes No  No  Yes Yes No  Yes Yes Yes Yes Yes Yes No  No  Yes Yes
 [91] Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes Yes Yes Yes Yes Yes Yes Yes
[109] Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes No  No
[127] No  Yes Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes No
[145] Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes
[163] Yes Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes
[181] Yes Yes Yes Yes Yes Yes No  Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes
[199] Yes Yes Yes Yes Yes Yes
Levels: No Yes
> |
```

```
> confusionMatrix(data= lgr_predictions, reference= test_data$heart_disease)
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  15  20
       Yes 84  85

               Accuracy : 0.4902
                 95% CI : (0.4197, 0.561)
    No Information Rate : 0.5147
    P-Value [Acc > NIR] : 0.7795

                  Kappa : -0.0397

 Mcnemar's Test P-Value : 6.506e-10

            Sensitivity : 0.15152
            Specificity : 0.80952
         Pos Pred Value : 0.42857
         Neg Pred Value : 0.50296
             Prevalence : 0.48529
         Detection Rate : 0.07353
   Detection Prevalence : 0.17157
      Balanced Accuracy : 0.48052

       'Positive' Class : No

> |
```

The above confusion matrix provides an evaluation of the logistic regression model's performance. The model's accuracy is low, and it seems to have difficulty distinguishing between the two classes, as indicated by the low sensitivity and relatively high specificity. The negative

kappa value suggests that the model is performing worse than chance, and Mcnemar's test indicates a significant difference between the predictions for the two classes. Overall, this model may not be well-suited for predicting heart disease based on elevated fasting blood sugar.

Next, the random forest model was examined:

```
> print(rf_model)
Random Forest

821 samples
  1 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 739, 739, 739, 739, 739, 738, ...
Resampling results:

  Accuracy  Kappa
  0.532251  0.04708533

Tuning parameter 'mtry' was held constant at a value of 2
```

```
44
45  rf_model<- train(heart_disease ~ fasting_blood_sugar, data = train_data, met
46  print(rf_model)
47
48  rf_predictions <- predict(rf_model, newdata = test_data)
49  print(rf_predictions)
50
```

50:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 · ~/geom practice/

```
> rf_predictions <- predict(rf_model, newdata = test_data)
> print(rf_predictions)
  [1] Yes Yes Yes No  Yes Yes No  Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes
 [19] No  Yes Yes Yes No  Yes Yes Yes Yes No  No  Yes Yes Yes Yes Yes No  Yes
 [37] No  Yes No  Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
 [55] Yes No  Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes
 [73] Yes Yes Yes No  No  Yes Yes No  Yes Yes Yes Yes Yes Yes No  No  Yes Yes
 [91] Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes Yes Yes Yes Yes Yes Yes Yes
[109] Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes No  No  Yes Yes No  No
[127] No  Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes Yes No
[145] Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes
[163] Yes Yes Yes Yes Yes Yes No  Yes Yes Yes Yes Yes Yes Yes No  Yes Yes Yes
[181] Yes Yes Yes Yes Yes Yes No  Yes No  Yes Yes Yes Yes Yes Yes Yes Yes Yes
[199] Yes Yes Yes Yes Yes Yes
Levels: No Yes
>
```

```
> confusionMatrix(data= rf_predictions, reference= test_data$heart_disease)
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  15  20
       Yes 84  85

               Accuracy : 0.4902
                 95% CI : (0.4197, 0.561)
    No Information Rate : 0.5147
    P-Value [Acc > NIR] : 0.7795

                  Kappa : -0.0397

 Mcnemar's Test P-Value : 6.506e-10

            Sensitivity : 0.15152
            Specificity : 0.80952
         Pos Pred Value : 0.42857
         Neg Pred Value : 0.50296
             Prevalence : 0.48529
         Detection Rate : 0.07353
   Detection Prevalence : 0.17157
      Balanced Accuracy : 0.48052

       'Positive' Class : No

> |
```

The random forest model had the same issues as the logistic regression model. Like the first model, the model's accuracy is low, and it also had difficulty distinguishing between the two classes, with low sensitivity and high specificity. The model also had a negative kappa value, which suggests that the model is performing worse than chance. Mcnemar's test again indicated a significant difference between the predictions for elevated fasting blood sugar and heart disease. In summary, both models were performing worse than chance and could not accurately predict heart disease from elevated fasting blood sugar from this data set.

Conclusion

Overall, based on the Heart Disease data set, there is no relationship between elevated glucose and heart disease, even though extensive research has solidified the link. The null hypothesis cannot be rejected, and the alternate hypothesis cannot be accepted. There are several possible explanations as to why this is the case. There may be insufficient data in the data set, which led to a poor relationship between the two variables. Since the data was from 1988 and

only had data from Europe and North America, it is possible that it did not reflect current global health trends and research. More diverse, current data needs to be analyzed to better determine the correlation between blood glucose and heart disease. As it stands based off the data used from the Heart Disease data set, it cannot be predicted if an individual will develop heart disease from blood glucose alone.

References

Bancks, M P., Ning, H., Allen,N B., Bertoni, A G., Carnethon, M R., Correa, A., Echouffo-Tcheugui, J B., Lange, L A., Lloyd-Jones, D M., Wilkins J T; (2019) Long-term Absolute Risk for Cardiovascular Disease Stratified by Fasting Glucose Level. *Diabetes Care* , 42 (3): 457–465. https://doi.org/10.2337/dc18-1773

Eckel, R. H., Bornfeldt, K. E., & Goldberg, I. J. (2021). Cardiovascular disease in diabetes, beyond glucose. Cell metabolism, 33(8), 1519–1545. https://doi.org/10.1016/j.cmet.2021.07.001

Lind, V., Hammar, N., Lundman, P. et al. (2021) Impaired fasting glucose: a risk factor for atrial fibrillation and heart failure. *Cardiovasc Diabetol* 20, 227. https://doi.org/10.1186/s12933-021-01422-3

Mashayekhi-Sardoo, H., Atkin, S. L., Montecucco, F., & Sahebkar, A. (2021). Potential Alteration of Statin-Related Pharmacological Features in Diabetes Mellitus. BioMed Research International, 1–9. https://doi-org.pitt.idm.oclc.org/10.1155/2021/6698743

Poznyak, A. V., Litvinova, L., Poggio, P., Sukhorukov, V. N., & Orekhov, A. N. (2022). Effect of Glucose Levels on Cardiovascular Risk. *Cells*, *11*(19), 3034. https://doi.org/10.3390/cells11193034

Project Code

```
library(readxl)
hd_data <-read_excel(file.choose("heart.xlsx"))
library(ggplot2)
library(caret)
heart_disease <- hd_data$target
fasting_blood_sugar <- hd_data$fbs
contingency_table <- table(heart_disease, fasting_blood_sugar)
print(contingency_table)
data <- data.frame(heart_disease = factor(heart_disease, levels = c(0, 1), labels = c("No", "Yes")),
                   fasting_blood_sugar = factor(fasting_blood_sugar, levels = c(0, 1), labels = c("No", "Yes")))
Scattered_box_plot <- ggplot(data, aes(x = heart_disease, fill = fasting_blood_sugar)) +
  geom_bar(position = "fill") +
  labs(x = "Heart Disease", y = "Proportion", fill = "Fasting Blood Sugar") +
  scale_fill_manual(values = c("No" = "lightblue", "Yes" = "lightpink")) +
  theme_minimal()
  print(Scattered_box_plot)

result <- chisq.test(table(heart_disease, fasting_blood_sugar))
print(result)

set.seed(123)

target_column <- 1
target <- data[, target_column]

predictor_columns <- 2:ncol(data)
```

```
predictors <- data[, predictor_columns]

train_indices <- createDataPartition(y = target, p = 0.8, list = FALSE)
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]
print(train_data)
print(test_data)

fitControl <- trainControl(method = "cv", number = 10, savePredictions = TRUE)

lgr_model <- train(heart_disease ~ fasting_blood_sugar, data = train_data, method =
"regLogistic", family = "binomial", trControl= fitControl, verbose= FALSE)
print(lgr_model)

lgr_predictions <- predict(lgr_model, newdata = test_data)
print(lgr_predictions)

confusionMatrix(data= lgr_predictions, reference= test_data$heart_disease)

rf_model<- train(heart_disease ~ fasting_blood_sugar, data = train_data, method = "rf", family =
"binomial", trControl= fitControl, verbose= FALSE)
print(rf_model)

rf_predictions <- predict(rf_model, newdata = test_data)
print(rf_predictions)

confusionMatrix(data= rf_predictions, reference= test_data$heart_disease)
```