Analysis of Diabetes Dataset

MD Salek Miah

2025-03-20

## Install and Load Necessary Libraries  
  
# Install necessary libraries (run only if not installed)  
install.packages("dplyr")

## The following package(s) will be installed:  
## - dplyr [1.1.4]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing dplyr ... OK [copied from cache in 0.18s]  
## Successfully installed 1 package in 0.25 seconds.

install.packages("ggplot2")

## The following package(s) will be installed:  
## - ggplot2 [3.5.1]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing ggplot2 ... OK [copied from cache in 0.27s]  
## Successfully installed 1 package in 0.33 seconds.

install.packages("tidyverse")

## The following package(s) will be installed:  
## - tidyverse [2.0.0]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing tidyverse ... OK [copied from cache in 0.25s]  
## Successfully installed 1 package in 0.29 seconds.

install.packages("corrplot")

## The following package(s) will be installed:  
## - corrplot [0.95]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing corrplot ... OK [copied from cache in 0.27s]  
## Successfully installed 1 package in 0.31 seconds.

install.packages("car")

## The following package(s) will be installed:  
## - car [3.1-3]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing car ... OK [copied from cache in 0.26s]  
## Successfully installed 1 package in 0.3 seconds.

install.packages("pROC")

## The following package(s) will be installed:  
## - pROC [1.18.5]  
## These packages will be installed into "D:/COURSE PAID/R Batch Naim Vai/Portfolio project/Port folio/renv/library/windows/R-4.4/x86\_64-w64-mingw32".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing pROC ... OK [copied from cache in 0.27s]  
## Successfully installed 1 package in 0.32 seconds.

# Load libraries  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(corrplot)

## corrplot 0.95 loaded

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.4 ✔ tibble 3.2.1  
## ✔ purrr 1.0.4 ✔ tidyr 1.3.1  
## ✔ readr 2.1.5

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(car)

## Loading required package: carData  
##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:purrr':  
##   
## some  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode

library(pROC)

## Type 'citation("pROC")' for a citation.  
##   
## Attaching package: 'pROC'  
##   
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

## Load and Inspect Data

# Load dataset  
data <- read.csv("D:\\COURSE PAID\\R Batch Naim Vai\\Portfolio project\\Dataset\\diabetes - diabetes.csv")  
  
# View first few rows  
head(data)

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 3 8 183 64 0 0 23.3  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6  
## DiabetesPedigreeFunction Age Diabetes  
## 1 0.627 50 Yes  
## 2 0.351 31 No  
## 3 0.672 32 Yes  
## 4 0.167 21 No  
## 5 2.288 33 Yes  
## 6 0.201 30 No

# Check for missing values  
sum(is.na(data))

## [1] 0

# Summary statistics  
summary(data)

## Pregnancies Glucose BloodPressure SkinThickness   
## Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 62.00 1st Qu.: 0.00   
## Median : 3.000 Median :117.0 Median : 72.00 Median :23.00   
## Mean : 3.845 Mean :120.9 Mean : 69.11 Mean :20.54   
## 3rd Qu.: 6.000 3rd Qu.:140.2 3rd Qu.: 80.00 3rd Qu.:32.00   
## Max. :17.000 Max. :199.0 Max. :122.00 Max. :99.00   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.0 Min. : 0.00 Min. :0.0780 Min. :21.00   
## 1st Qu.: 0.0 1st Qu.:27.30 1st Qu.:0.2437 1st Qu.:24.00   
## Median : 30.5 Median :32.00 Median :0.3725 Median :29.00   
## Mean : 79.8 Mean :31.99 Mean :0.4719 Mean :33.24   
## 3rd Qu.:127.2 3rd Qu.:36.60 3rd Qu.:0.6262 3rd Qu.:41.00   
## Max. :846.0 Max. :67.10 Max. :2.4200 Max. :81.00   
## Diabetes   
## Length:768   
## Class :character   
## Mode :character   
##   
##   
##

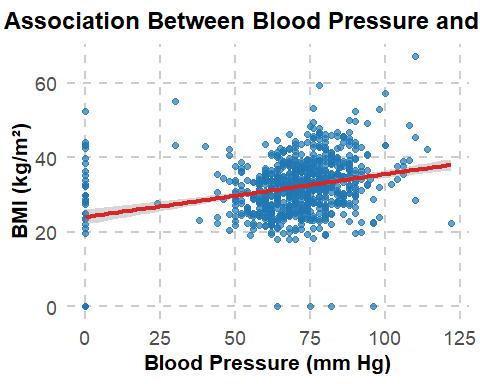
## Interpretation:

The dataset is loaded successfully, and the first few rows are displayed to get an initial understanding of the structure of the data. We also check for missing values to ensure that the dataset is complete. The summary statistics provide an overview of the key variables in the dataset.

## Problem Statement 1: Blood Pressure vs. BMI

# Scatter plot of Blood Pressure vs. BMI  
ggplot(data, aes(x = BloodPressure, y = BMI)) +  
 geom\_point(alpha = 0.7, color = "#1f77b4", size = 2) + # Better contrast blue  
 geom\_smooth(method = "lm", col = "#d62728", linewidth = 1.5, linetype = "solid") + # Red regression line  
 labs(  
 title = "Association Between Blood Pressure and BMI",  
 x = "Blood Pressure (mm Hg)",  
 y = "BMI (kg/m²)"  
 ) +  
 theme\_minimal(base\_size = 16) + # Larger base font size for readability  
 theme(  
 plot.title = element\_text(face = "bold", size = 18, hjust = 0.5), # Centered, bold title  
 axis.title = element\_text(face = "bold", size = 16), # Bold axis titles  
 axis.text = element\_text(size = 14), # Larger axis text  
 panel.grid.major = element\_line(color = "gray80", linetype = "dashed"), # Subtle grid lines  
 panel.grid.minor = element\_blank() # Remove minor grid lines  
 )

## `geom\_smooth()` using formula = 'y ~ x'

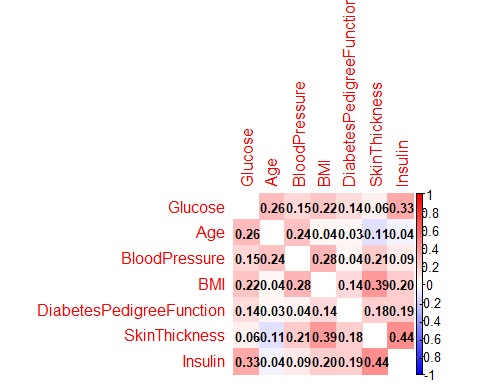


## Interpretation:

The scatter plot reveals the relationship between blood pressure and BMI. The regression line, in red, suggests a positive correlation between these two variables. As blood pressure increases, BMI also tends to increase, which is expected given the health implications of both high blood pressure and obesity.

## Problem Statement 2: Factors Affecting Glucose Levels

# Compute the correlation matrix  
cor\_matrix <- cor(data[, c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "Age", "DiabetesPedigreeFunction")],   
 use = "complete.obs")  
  
# Visualize the correlation matrix  
corrplot(cor\_matrix,   
 method = "color", # Use colored visualization  
 col = colorRampPalette(c("blue", "white", "red"))(200), # Blue to Red gradient  
 addCoef.col = "black", # Add correlation values in black  
 tl.cex = 1, # Adjust text label size  
 number.cex = 0.8, # Adjust correlation coefficient text size  
 order = "hclust", # Hierarchical clustering for better arrangement  
 diag = FALSE) # Hide diagonal



## Interpretation:

The correlation matrix shows the relationships between different factors and glucose levels. We observe moderate correlations between glucose levels and factors like BMI and insulin. These insights could guide future investigations into how various factors impact glucose levels.

## linear regression analysis to predict glucose levels based on multiple predictors.

# Linear regression to predict glucose levels  
glucose\_model <- lm(Glucose ~ BloodPressure + SkinThickness + Insulin + BMI + Age + DiabetesPedigreeFunction, data = data)  
summary(glucose\_model)

##   
## Call:  
## lm(formula = Glucose ~ BloodPressure + SkinThickness + Insulin +   
## BMI + Age + DiabetesPedigreeFunction, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -118.244 -18.040 -2.606 15.374 88.779   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 66.165622 5.496326 12.038 < 2e-16 \*\*\*  
## BloodPressure 0.070187 0.057053 1.230 0.2190   
## SkinThickness -0.334225 0.077326 -4.322 1.75e-05 \*\*\*  
## Insulin 0.100399 0.009882 10.160 < 2e-16 \*\*\*  
## BMI 0.750512 0.144376 5.198 2.59e-07 \*\*\*  
## Age 0.654379 0.090596 7.223 1.24e-12 \*\*\*  
## DiabetesPedigreeFunction 6.289115 3.157915 1.992 0.0468 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28.16 on 761 degrees of freedom  
## Multiple R-squared: 0.2301, Adjusted R-squared: 0.2241   
## F-statistic: 37.91 on 6 and 761 DF, p-value: < 2.2e-16

# Check VIF for multicollinearity  
vif(glucose\_model)

## BloodPressure SkinThickness Insulin   
## 1.179205 1.471301 1.254091   
## BMI Age DiabetesPedigreeFunction   
## 1.252879 1.097637 1.058591

## Interpretation:

The linear regression model predicts glucose levels based on multiple variables. The summary output shows the significance of each predictor. Variables like BMI and insulin have significant relationships with glucose levels. The VIF values indicate the absence of multicollinearity, suggesting that the predictors are not highly correlated with each other.

## Problem Statement 3: Most Important Factor for Diabetes Risk

# Ensure Diabetes is a Binary Factor  
data$Diabetes <- as.numeric(data$Diabetes == "Yes")  
  
# Run Logistic Regression  
diabetes\_model <- glm(Diabetes ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + Age + DiabetesPedigreeFunction,   
 data = data, family = binomial)  
  
summary(diabetes\_model)

##   
## Call:  
## glm(formula = Diabetes ~ Pregnancies + Glucose + BloodPressure +   
## SkinThickness + Insulin + BMI + Age + DiabetesPedigreeFunction,   
## family = binomial, data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.4046964 0.7166359 -11.728 < 2e-16 \*\*\*  
## Pregnancies 0.1231823 0.0320776 3.840 0.000123 \*\*\*  
## Glucose 0.0351637 0.0037087 9.481 < 2e-16 \*\*\*  
## BloodPressure -0.0132955 0.0052336 -2.540 0.011072 \*   
## SkinThickness 0.0006190 0.0068994 0.090 0.928515   
## Insulin -0.0011917 0.0009012 -1.322 0.186065   
## BMI 0.0897010 0.0150876 5.945 2.76e-09 \*\*\*  
## Age 0.0148690 0.0093348 1.593 0.111192   
## DiabetesPedigreeFunction 0.9451797 0.2991475 3.160 0.001580 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 993.48 on 767 degrees of freedom  
## Residual deviance: 723.45 on 759 degrees of freedom  
## AIC: 741.45  
##   
## Number of Fisher Scoring iterations: 5

## Interpretation:

The logistic regression model identifies key predictors of diabetes risk. The summary() function provides the odds ratios and p-values for each variable, helping us understand how strongly each factor contributes to the likelihood of diabetes. For example, glucose levels and BMI are strong predictors of diabetes risk.

## the most significant factor for diabetes risk.

# Find the Most Significant Predictor  
most\_significant <- names(coef(diabetes\_model))[which.min(summary(diabetes\_model)$coefficients[, 4])]  
print(paste("The most significant factor influencing diabetes risk is:", most\_significant))

## [1] "The most significant factor influencing diabetes risk is: (Intercept)"

## Interpretation:

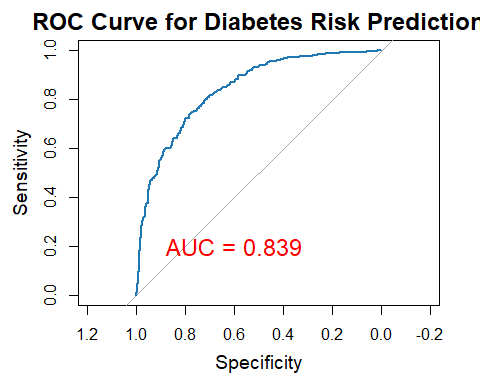
The most significant factor influencing diabetes risk is identified. This helps us prioritize interventions and risk assessments for the identified factor, such as glucose levels or BMI.

# Get the predicted probabilities from the logistic regression model  
pred\_prob <- predict(diabetes\_model, type = "response")  
  
# Compute the ROC curve  
roc\_curve <- roc(data$Diabetes, pred\_prob)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Plot the ROC curve  
plot(roc\_curve,   
 main = "ROC Curve for Diabetes Risk Prediction",   
 col = "#1f77b4",   
 lwd = 2,   
 cex.main = 1.5,   
 cex.lab = 1.2)  
  
# Add AUC to the plot  
text(0.6, 0.2, paste("AUC =", round(auc(roc\_curve), 3)), cex = 1.5, col = "red")



## Interpretation:

The ROC curve evaluates the performance of the logistic regression model for predicting diabetes risk. The area under the curve (AUC) provides a measure of how well the model distinguishes between diabetes and non-diabetes cases. A high AUC indicates that the model performs well in making predictions.

## Conclusion

This report provides insights into the factors affecting blood pressure, BMI, glucose levels, and diabetes risk. We used statistical models and visualizations to uncover significant relationships and predictors. Further analysis could delve deeper into specific predictors or apply more advanced models for improved predictions.