Experiment 7

April 21, 2025

[ ]:

```r
# Load required libraries
set.seed(1)
library(tidyverse)
library(caret)
library(glmnet)
library(mlbench)
library(randomForest)
```

[ ]:

```r
# Load Pima Indians Diabetes dataset
data("PimaIndiansDiabetes2")
df <- PimaIndiansDiabetes2
```

[ ]:

```r
# Check structure & missing values
glimpse(df)
summary(df)
df <- na.omit(df)
```

[ ]:

```r
preProc <- preProcess(df[, -9], method = c("center", "scale"))
df_scaled <- predict(preProc, df)
```

[ ]:

```r
df_scaled <- df_scaled %>% mutate(bmi_age_ratio = mass / age)
```

[ ]:

```r
cor_matrix <- cor(df_scaled %>% select(-diabetes))
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.7)
```

[ ]:

```r
ctrl <- rfeControl(functions = rfFuncs, method = "cv", number = 10)
rfe_result <- rfe(
  x = df_scaled %>% select(-diabetes),
  y = df_scaled$diabetes,
  sizes = 1:8, # Test subsets of 1 to 8 features
  rfeControl = ctrl
)

# Top selected features
print(rfe_result)
plot(rfe_result, type = c("g", "o"))
```

[ ]:

```
x <- model.matrix(diabetes ~ ., df_scaled)[, -1] # Exclude intercept
y <- ifelse(df_scaled$diabetes == "pos", 1, 0)

# Fit LASSO
cv_lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial")
plot(cv_lasso)

# Coefficients at optimal lambda
coef(cv_lasso, s = "lambda.min")
```

[ ]:

```
trainIndex <- createDataPartition(df_scaled$diabetes, p = 0.8, list = FALSE)
train <- df_scaled[trainIndex, ]
test <- df_scaled[-trainIndex, ]
```

[ ]:

```
model_all <- train(
  diabetes ~ .,
  data = train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 10)
)

pred_all <- predict(model_all, test)
confusionMatrix(pred_all, test$diabetes)
```

[ ]:

```
model_selected <- train(
  diabetes ~ glucose + mass + bmi_age_ratio,
  data = train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 10)
)

# Predictions
pred_selected <- predict(model_selected, test)
confusionMatrix(pred_selected, test$diabetes)
```

[ ]:

```

```