weekly assignment 04

Xiang Li

2024/3/4

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
data_df = read.csv("data.csv")
data_df = data_df[, -c(1, 33)]
```

1

```
set.seed(519)
n = nrow(data_df)
train_id = sample(n, round(n * 0.8))
train_df = data_df[train_id, ]
test_df = data_df[-train_id, ]
```

2

Choose 3 methods: QDA, Logistic Regression, KNN with k=4.

3

```
train_control = trainControl(method = "cv", number = 10)
set.seed(519)
qda_model = train(diagnosis ~ ., method = "qda", trControl = train_control, data = train_df)
qda_model$results
```

```
##   parameter  Accuracy     Kappa AccuracySD    KappaSD
## 1      none 0.9517391 0.8972847 0.03558044 0.07514161
```

```r
set.seed(519)
lr_model = train(diagnosis ~ ., method = "glm", trControl = train_control, data = train_df)
lr_model$results
```

```
##   parameter Accuracy    Kappa AccuracySD    KappaSD
## 1      none 0.934058 0.861256 0.04530558 0.09237242
```

```r
set.seed(519)
knn_model = train(diagnosis ~ ., method = "knn", trControl = train_control, data = train_df,
    tuneGrid = data.frame(k = 4), preProcess = c("center", "scale"))
knn_model$results
```

```
##   k  Accuracy     Kappa AccuracySD    KappaSD
## 1 4 0.9671014 0.9284466 0.02566855 0.05610628
```

Based on the accuracy of cross-validation, the most accurate model is KNN with k=4.

4

```r
set.seed(519)
knn_model_fin = train(diagnosis ~ ., method = "knn", trControl = trainControl(method = "none"),
    data = train_df, tuneGrid = data.frame(k = 4), preProcess = c("center", "scale"))
knn_pre = predict(knn_model_fin, newdata = test_df)
confusionMatrix(knn_pre, as.factor(test_df$diagnosis))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 73  5
##          M  1 35
##
##                Accuracy : 0.9474
##                  95% CI : (0.889, 0.9804)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : 2.978e-14
##
##                   Kappa : 0.8817
##
##  Mcnemar's Test P-Value : 0.2207
##
##             Sensitivity : 0.9865
##             Specificity : 0.8750
##          Pos Pred Value : 0.9359
##          Neg Pred Value : 0.9722
##              Prevalence : 0.6491
##          Detection Rate : 0.6404
##    Detection Prevalence : 0.6842
##       Balanced Accuracy : 0.9307
```

```
##
##        'Positive' Class : B
##
```

On the test set, accuracy of the selected model is 0.9497, specificity is 0.8750, and sensitivity is 0.9865.