Classificação de Diabetes

Luan Roger Santos Santana

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

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

Preparing data

Loading Data

```
data_raw <- read.csv("../data_sets/Material 03 - 9 - C - Diabetes - Dados.csv")
data_raw_new_cases <- read.csv("../data_sets/Material 03 - 9 - C - Diabetes - Dados - Novos Casos.csv")
print(head(data_raw))
```

```
##   num preg0nt glucose pressure triceps insulin mass pedigree age diabetes
## 1   1       6     148       72      35       0 33.6    0.627  50      pos
## 2   2       1      85       66      29       0 26.6    0.351  31      neg
## 3   3       8     183       64       0       0 23.3    0.672  32      pos
## 4   4       1      89       66      23      94 28.1    0.167  21      neg
## 5   5       0     137       40      35     168 43.1    2.288  33      pos
## 6   6       5     116       74       0       0 25.6    0.201  30      neg
```

```
print(head(data_raw_new_cases))
```

```
##   num preg0nt glucose pressure triceps insulin mass pedigree age diabetes
## 1   1       7     130       72      37       0 33.6    0.980  50        ?
## 2   2       2      81       66      29       0 32.6    0.351  31        ?
## 3   3       5      23       64       0       0 23.3    0.672  15        ?
```

Cleaning data

```
data <- data_raw[,!(names(data_raw) %in% c('num'))]
data_new_cases <- data_raw_new_cases[,!(names(data_raw_new_cases) %in% c('num'))]
print(head(data))
```

```
##   preg0nt glucose pressure triceps insulin mass pedigree age diabetes
## 1       6     148       72      35       0 33.6    0.627  50      pos
## 2       1      85       66      29       0 26.6    0.351  31      neg
## 3       8     183       64       0       0 23.3    0.672  32      pos
## 4       1      89       66      23      94 28.1    0.167  21      neg
## 5       0     137       40      35     168 43.1    2.288  33      pos
## 6       5     116       74       0       0 25.6    0.201  30      neg
```

```
print(head(data_new_cases))
```

```
##   preg0nt glucose pressure triceps insulin mass pedigree age diabetes
## 1       7     130       72      37       0 33.6    0.980  50        ?
## 2       2      81       66      29       0 32.6    0.351  31        ?
```

```
## 3       5      23       64       0       0 23.3    0.672  15        ?
```

Creating data partitioning

```
set.seed(1988)
ran <- sample(1:nrow(data), 0.8 * nrow(data))
training_data <- data[ran,]
test_data <- data[-ran,]
```

Training

Using KNN

Creating the model

```
tuneGrid <- expand.grid(k = c(1,3,5,7,9))
set.seed(1988)
knn <- train(diabetes ~ ., data = training_data, method = "knn", tuneGrid=tuneGrid)
print(knn)
```

```
## k-Nearest Neighbors
##
## 614 samples
##   8 predictor
##   2 classes: 'neg', 'pos'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 614, 614, 614, 614, 614, 614, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   1  0.6742899  0.2869675
##   3  0.6728829  0.2832179
##   5  0.6917942  0.3186256
##   7  0.6947715  0.3227047
##   9  0.7057915  0.3431622
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

```
prediction.knn <- predict(knn, test_data)
cf_matrix <- confusionMatrix(prediction.knn, as.factor(test_data$diabetes))
print(cf_matrix)
```

Checking the model with training data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction neg pos
##        neg  87  24
##        pos  16  27
```

```
##
##                Accuracy : 0.7403
##                  95% CI : (0.6635, 0.8075)
##     No Information Rate : 0.6688
##     P-Value [Acc > NIR] : 0.03415
##
##                   Kappa : 0.3895
##
##  Mcnemar's Test P-Value : 0.26838
##
##             Sensitivity : 0.8447
##             Specificity : 0.5294
##          Pos Pred Value : 0.7838
##          Neg Pred Value : 0.6279
##              Prevalence : 0.6688
##          Detection Rate : 0.5649
##    Detection Prevalence : 0.7208
##       Balanced Accuracy : 0.6870
##
##        'Positive' Class : neg
##
```

Checking for new cases

```
prediction.knn_new_data <- predict(knn, data_new_cases)
data_new_cases$diabetes <- NULL
result <- cbind(data_new_cases, diabetes=prediction.knn_new_data)
print(result)
```

```
##   preg0nt glucose pressure triceps insulin mass pedigree age diabetes
## 1       7     130       72      37       0 33.6    0.980  50      pos
## 2       2      81       66      29       0 32.6    0.351  31      neg
## 3       5      23       64       0       0 23.3    0.672  15      neg
```