Classificação de Imposto de Renda

Luan Roger Santos Santana

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

Preparing data

Loading Data

```
data_raw <- read.csv("../data_sets/Material 03 - 7 - C - IR - Dados.csv")
data_raw_new_cases <- read.csv("../data_sets/Material 03 - 7 - C - IR - Dados - Novos Casos.csv")
```

Cleaning data

```
data <- data_raw
data_new_cases <- data_raw_new_cases
print(head(data))
```

```
##   rest     ecivil rendimento sonegador
## 1  Sim   Solteiro     125000       Sim
## 2  Nao     Casado     100000       Nao
## 3  Nao   Solteiro      70000    Talvez
## 4  Sim     Casado     120000       Sim
## 5  Nao Divorciado      95000    Talvez
## 6  Nao     Casado      60000       Nao
```

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

```
##   rest   ecivil rendimento sonegador
## 1  Sim Solteiro      99000         ?
## 2  Nao   Casado       9999         ?
## 3  Nao Solteiro      73200         ?
```

Creating data partitioning

```
set.seed(1988)
ran <- sample(1:nrow(data), 0.8 * nrow(data))
ran <- createDataPartition(data$sonegador, p = 0.80, list = F)
training_data <- data[ran,]
test_data <- data[-ran,]
```

Training

Using KNN

Creating the model

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

```
## k-Nearest Neighbors
##
## 40 samples
##  3 predictor
##  3 classes: 'Nao', 'Sim', 'Talvez'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 40, 40, 40, 40, 40, 40, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   1  0.9653575  0.9463258
##   3  0.7895162  0.6853010
##   5  0.5333563  0.3377246
##   7  0.5128193  0.3074524
##   9  0.4570315  0.2254728
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 1.
```

```r

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

Checking the model with training data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Nao Sim Talvez
##     Nao      4   0      0
##     Sim      0   3      0
##     Talvez   0   0      3
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.6915, 1)
##     No Information Rate : 0.4
##     P-Value [Acc > NIR] : 0.0001049
##
##                   Kappa : 1
```

```
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Nao Class: Sim Class: Talvez
## Sensitivity                 1.0        1.0           1.0
## Specificity                 1.0        1.0           1.0
## Pos Pred Value              1.0        1.0           1.0
## Neg Pred Value              1.0        1.0           1.0
## Prevalence                  0.4        0.3           0.3
## Detection Rate              0.4        0.3           0.3
## Detection Prevalence        0.4        0.3           0.3
## Balanced Accuracy           1.0        1.0           1.0
```

Checking for new cases

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

```
##   rest   ecivil rendimento sonegador
## 1  Sim Solteiro      99000       Nao
## 2  Nao   Casado       9999       Nao
## 3  Nao Solteiro      73200       Nao
```