K-Nearest Neighbors (K-NN) Verkhovod

# Download the data

set.seed(123)  
setwd('C:/Users/VerkhovodTS/Desktop/R')  
f\_train <- read.csv2('clients\_train.csv', header = TRUE, encoding = 'UNICOD')  
f\_test <- read.csv2('clients\_test.csv', header = TRUE, encoding = 'UNICOD')  
f\_train <- f\_train[-1]  
f\_test <- f\_test[-1]

# Fitting & predicting

library(class)  
y = knn(train = f\_train[,c('LOAN\_PRODUCT\_TYPE','LOAN.TERM')],  
 test = f\_test[,c('LOAN\_PRODUCT\_TYPE','LOAN.TERM')],  
 cl = f\_train[, 'DELAY'],  
 k = 20,  
 prob = TRUE)

Висновок: і навчання, і прогнозування за моделлю k найближчих сусідів здійснюється однією функцією. У результаті отримуємо вектор класів об’єктів.

# Confusion Matrix

cm = table(f\_test[, 'DELAY'], y == '1')  
print(cm)

##   
## FALSE TRUE  
## 0 190 149  
## 1 132 329

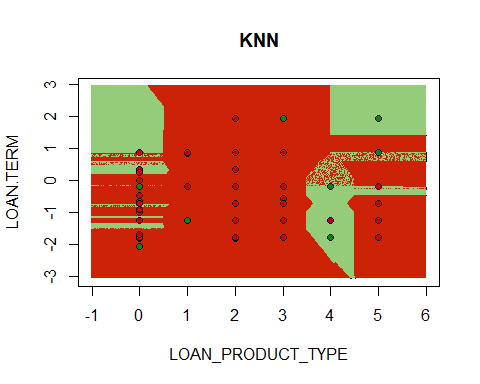
Висновок: точність моделі – (190+329) / 800 = 64,88 %, частка невірно класифікованих випадків – (132+149) / 800 = 35,13 %. Чутливість – 329 / (132+329) = 71,37 %, специфічність – 190 / (190+149) = 56,05 %, тобто модель більш чутлива до виявлення позитивних випадків (кредиторів, що мають заборгованість).

**Visualising the Test set results**

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

set = f\_test[,c('LOAN\_PRODUCT\_TYPE','LOAN.TERM','DELAY')]  
X1 = seq(min(set['LOAN\_PRODUCT\_TYPE']) - 1, max(set['LOAN\_PRODUCT\_TYPE']) + 1, by = 0.01)  
X2 = seq(min(set['LOAN.TERM']) - 1, max(set['LOAN.TERM']) + 1, by = 0.01)  
grid\_set = expand.grid(X1, X2)  
colnames(grid\_set) = c('LOAN\_PRODUCT\_TYPE','LOAN.TERM')  
y\_grid = knn(train = f\_train[,c('LOAN\_PRODUCT\_TYPE','LOAN.TERM')], test = grid\_set, cl = f\_train[, 'DELAY'], k = 5)  
plot(set[, -3],  
 main = 'KNN',  
 xlab = 'LOAN\_PRODUCT\_TYPE', ylab = 'LOAN.TERM',  
 xlim = range(X1), ylim = range(X2))  
contour(X1, X2, matrix(as.numeric(y\_grid), length(X1), length(X2)), add = TRUE)  
points(grid\_set, pch = '.', col = ifelse(y\_grid == 1, '#cc2208', '#94cc7a'))  
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'red3', 'green4'))



Висновок: на графіку червоним позначені випадки затримки з повернення кредиту, зеленим – хороші кредитори. Червоним виділена зона високої ймовірності неповернення кредиту. Модель описує нелінійний варіант розподіляючої кривої.