* **Название и номер задания:**

Задание ДР

* **Постановка задачи:**

1. Примените метод деревьев решений для задачи классификации (для использованных ранее или новых данных).

2. Если число признаков m = 2, визуализуйте данные (постройте облако точек).

3. Исследуйте дерево решений; если позволяет размерность, постройте его график.

4. Проанализируйте точность полученных решений для тестовых данных (с известным значением переменной отклика), сравните результаты с ранее полученными.

5. Задайте несколько новых данных, покажите соответствующие точки на графике (при m = 2) (выделите их другим цветом).

6. Определите класс для новых данных (т.е. для данных, где известны только значения факторов, но не переменной отклика).

* **Студент:**

Санамян Артак

* **Номер группы:**

09-715(1)

* **Ссылка на источник данных:**

[**https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/**](https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/)

* **Постановка задачи**

Для демонстрации маркетинговой модели RFMTC (модифицированная версия RFM) в исследовании была использована донорская база данных центра обслуживания переливания крови в городе Синь-Чу на Тайване. Центр передает автобус службы переливания крови в один университет в городе Синь-Чу, чтобы собирать кровь, которую сдают примерно каждые три месяца. Для построения модели FRMTC произвольно выбрали 748 доноров из базы данных доноров. Эти данные 748 доноров, каждый из которых включает Р – сколько прошло месяцев с момента последней сдачи крови, Ф – общее количество сдачи крови (раз), М – общее количество сданной крови, T – сколько месяцев прошло с момента первой сдачи крови и переменная отклика – сдача в определенный месяц (март 2017), 1 – кровь сдавалась, а 0 – означает, что кровь не сдавалась.

По исходной выборке необходимо реализовать метод построения деревьев решений.

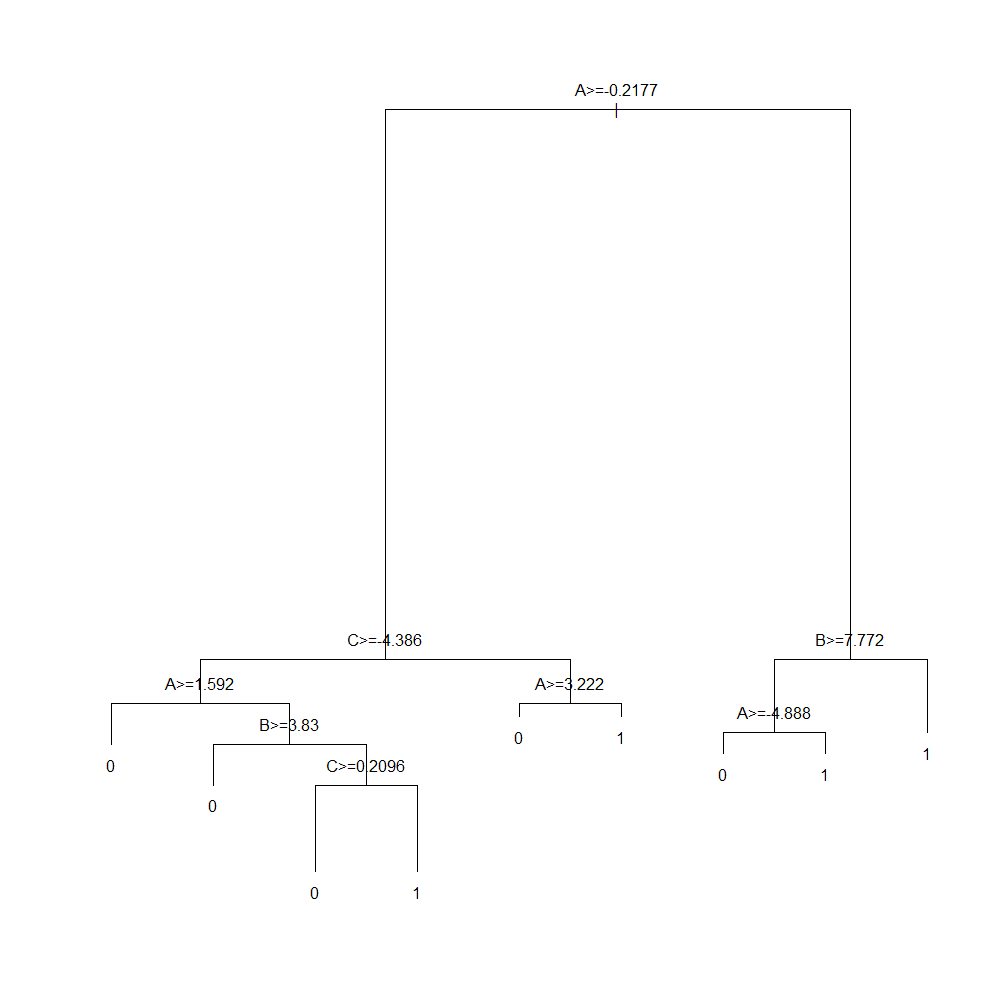
* **Описание выполнения работы**

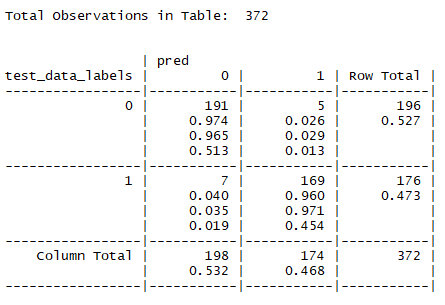
Разделим исходную выборку на обучающую и тестовую (необходимость этого заключается в том, что по построенной модели будет необходимо оценить истинность предсказанных значений).

Так как число атрибутов выборки больше двух, не представляется возможным привести графическую интерпретацию.

При помощи обучающей выборки, построим модель метода ДР.

Построим дерево решений:



Рассмотрим кросс-валидационую таблицу полученной модели:

Из полученной таблицы видно, что 12 из 372 значений классифицированы неверно, то есть метод ошибается на 0,03%, когда метод НБК не ошибся ни разу.

Также стоит заметить, что ситуация, когда фальшивая купюра классифицируется как подлинная, является более значимой, чем обратная ошибка. Данный метод классифицировал 5 фальшивых купюр как подлинные и 7 подлинных как ложные.

Сравним полученные целевые значения для тестовой выборки и истинных переменных отклика. Добавим десять случайных значений из тестовой выборки и сравним с предсказанными значениями:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

Исходя из полученных результатов, можно увидеть, что ожидаемые значения переменной отклика совпадают со значениями, вычисленными при помощи метода ДР.

* **Приложение**

rm (list=ls())

A=scan("aaa4.txt")

Data\_null=matrix(A,ncol=5,byrow=TRUE)

colnames(Data\_null)<-c('A','B','C','D','R')

dims = dim(Data\_null)

n = dims[1]

m = dims[2]

n\_data = 372

n\_train = 1000

set.seed(1983)

Data=Data\_null[order(runif(1372)),]

train\_data = Data[1:n\_train,]

train\_data\_labels = train\_data[,5]

train\_data\_labels<-factor(train\_data\_labels)

test\_data = Data[(n\_train+1):n, ]

test\_data\_labels = test\_data[,5]

library(rpart)

train\_data\_fr = as.data.frame.matrix(train\_data)

my\_tree <- rpart(R ~ A + B + C + D, method="class", data=train\_data\_fr)

plot(my\_tree)

text(my\_tree, cex = 1, all = FALSE, use.n = FALSE)

test\_data <- test\_data[,1:4]

test\_data <- data.frame(test\_data)

pred <- predict(my\_tree, test\_data, type="class")

library("gmodels")

CrossTable(x = test\_data\_labels, y = pred, prop.chisq=FALSE)