МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра вычислительной техники

**Лабораторная работа №3**

**по дисциплине «Методы анализа данных»**

**Применение деревьев классификации в решении задач интеллектуального анализа данных (ИАД) средствами интегрированной системы Statistica и языка R**

|  |  |
| --- | --- |
| Студент | Великжанин С.И. |
| Вариант | 3 |
| Группа | АВТ-010 |
| Преподаватель | Альсова О.К. |
| Дата сдачи | 11.12.2023 |

Новосибирск, 2023 г.

1. **Цели**

* изучить алгоритмы и методы деревьев классификации на примере решения задачи ИАД;
* исследовать эффективность использования алгоритмов и методов деревьев классификации для решения прикладной задачи;
* ознакомиться и получить практические навыки работы с модулями интегрированной системы *Statistica* и среды *R,* реализующими методы деревьев классификации.

1. **Постановка задачи ИАД**

Изучаются показатели работы программистов крупной организации.Необходимо выделить группы (классы) программистов в соответствии с рядом устойчивых признаков. Разделение программистов на группы проводится с целью установления и обоснования тарифных разрядов и размера заработной платы для каждой выделенной группы. Также требуется построить правило отнесения программиста к одной из выделенных групп (классов).

Исходные данные для проведения статистического анализа представлены в табл. 1. Рассматриваются следующие показатели (признаки) для каждого программиста:

* возраст;
* время написания первой тестовой программы, в час.;
* время написания второй тестовой программы, в час.;
* стаж работы по специальности в данной организации;
* образование (непрофильное – 0; профильное - 1);

наличие сертификатов о повышении квалификации в области программирования и программного обеспечения (0 –сертификаты отсутствуют; 1- сертификаты 1-го уровня; 2 - сертификаты 2-го уровня, 3-  сертификаты 3-го уровня).

**вариант 3.**

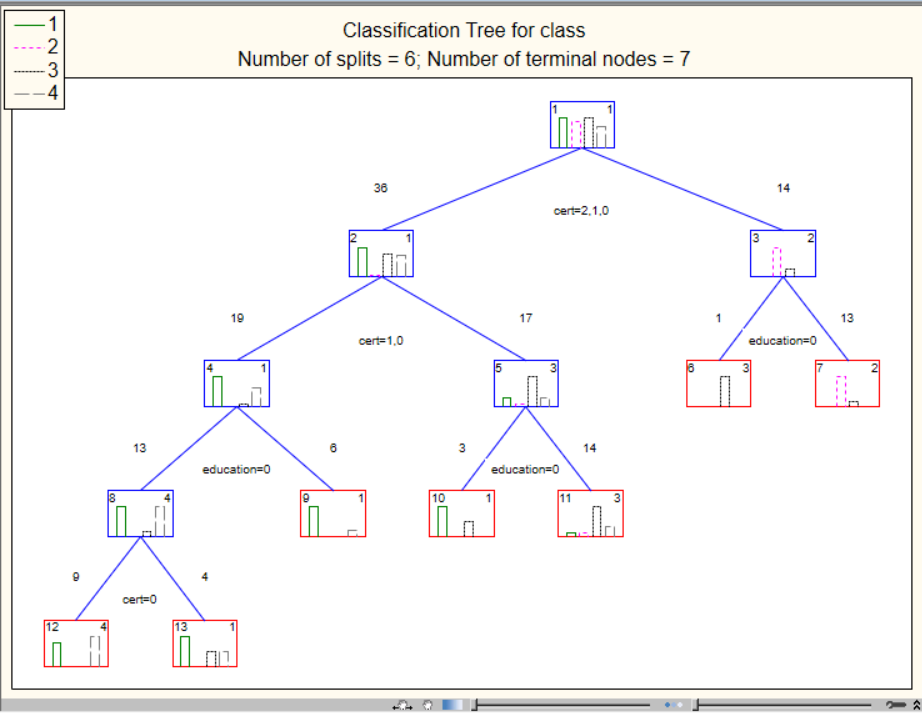
*Таблица 1.*

***Исходные данные (фрагмент)***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **номер** | **возраст** | **время   тест 1** | **время   тест 2** | **стаж** | **образование** | **сертификат** |
| 1 | 39 | 5,4 | 2,8 | 0,8 | 0 | 0 |
| 2 | 31 | 2,0 | 1,7 | 3,5 | 1 | 2 |
| 3 | 32 | 2,4 | 1,8 | 6,1 | 1 | 3 |
| 4 | 36 | 3,4 | 1,8 | 3,3 | 1 | 2 |
| 5 | 29 | 5,4 | 2,7 | 0,7 | 0 | 0 |
| ... | … | … | … | … | … | … |
| 50 | 39 | 5,4 | 2,6 | 0,7 | 1 | 0 |

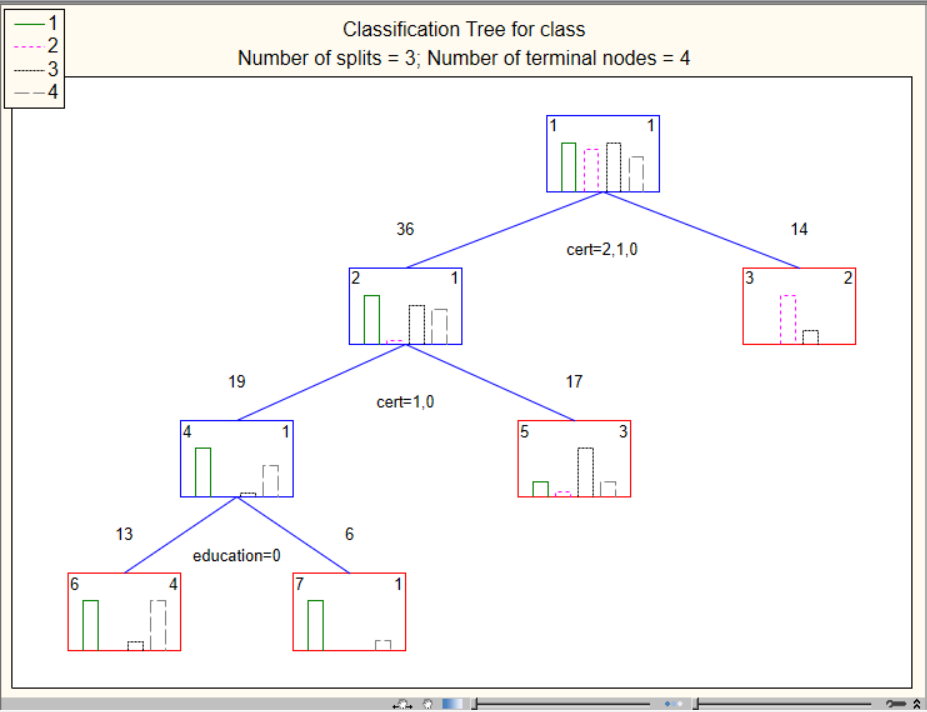
1. **Ход работы**

Построение деревьев классификации:

**

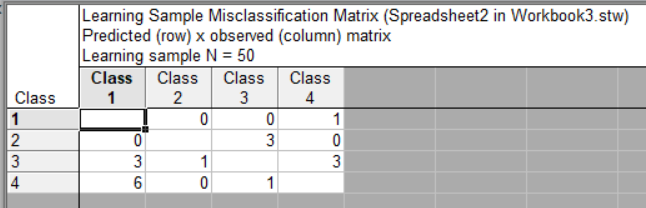
*Рис. 1 Дерево классификации на основе кол. и кач. Признаков*

Вывод: полученное дерево является несбалансированным, имеет высокую сложность.

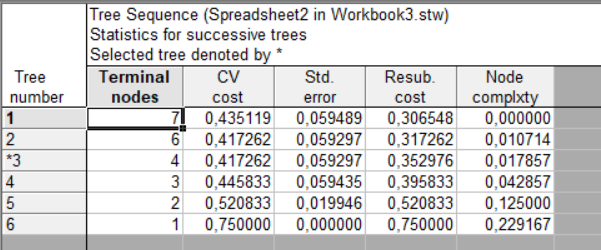
**

*Рис. 2 Дерево классификации, усеченное по минимальной цене сложности*

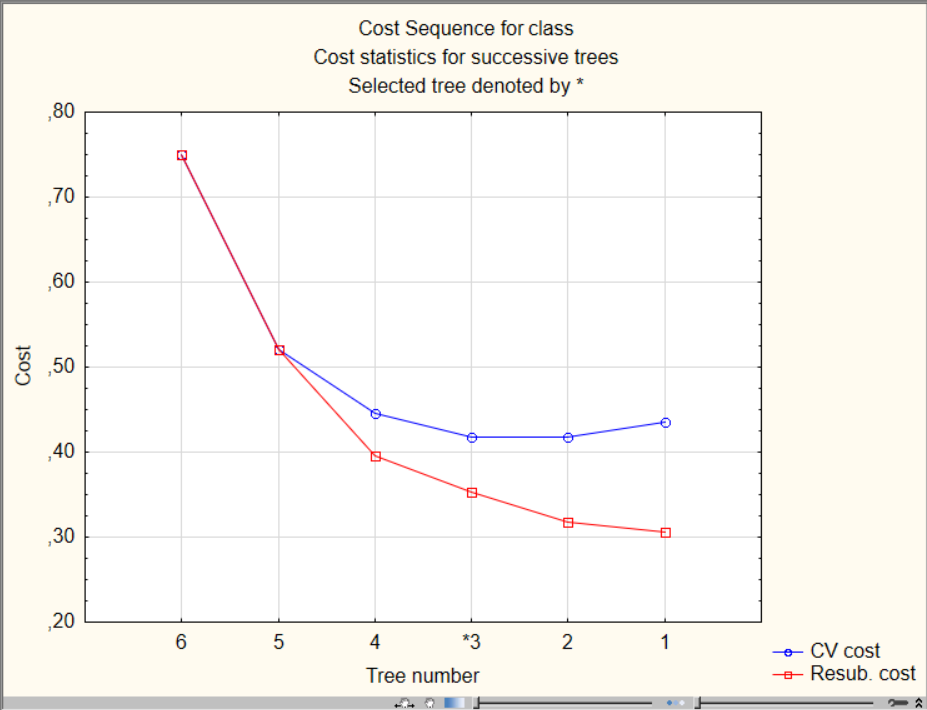
Вывод: Усечение дерева по минимальной цене сложности позволяет уменьшить сложность, ценой увеличения стоймости. Однако дерево по-прежнему является несбалансированным.

**

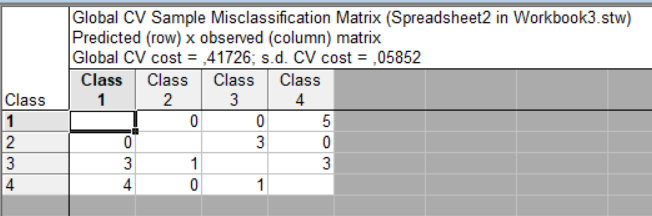
*Рис. 3 Матрица ошибок*

**

*Рис. 4 Последовательность оптимально усеченных деревьев*

**

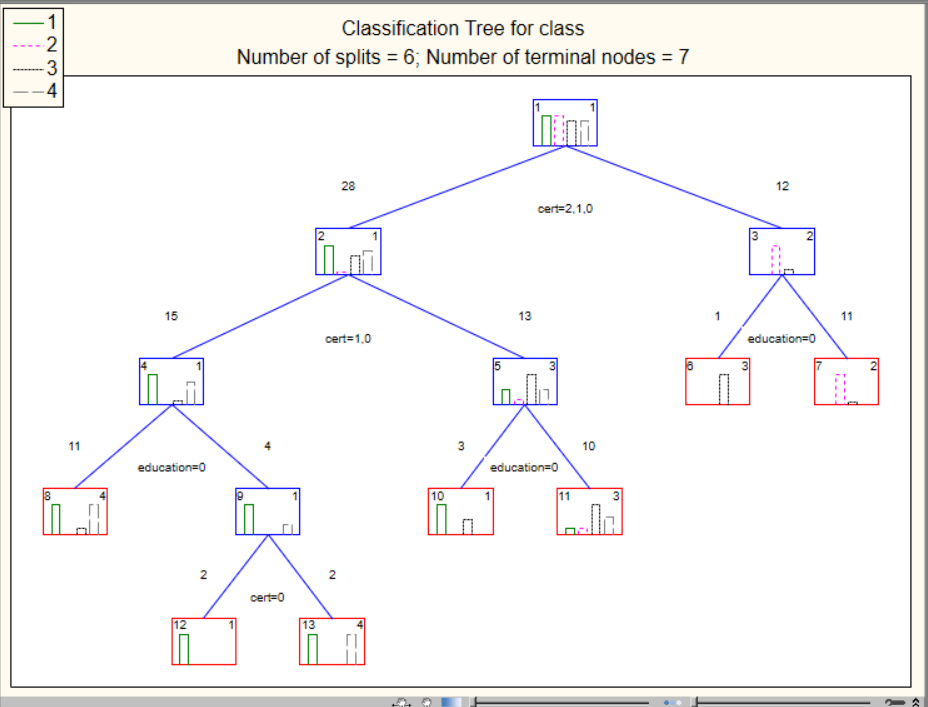
*Рис. 5 График цены обучения и цены кросс-проверки*

**

*Рис. 6 Матрица ошибок по результатам глобальной кросс-проверки*

Вывод: при смене правила останова растет цена проверки, но также уменьшается число ошибок при классификации.

Выделение тестовой выборки:

**

*Рис. 7 Дерево классификации с тестовой выборкой*

Вывод: выделение тестовой выборки позволяет лучше сбалансировать дерево, однако возрастает сложность и цена проверки.

Построение деревьев в R:

> split <- sample.split(data2$groups, SplitRatio = 0.8)

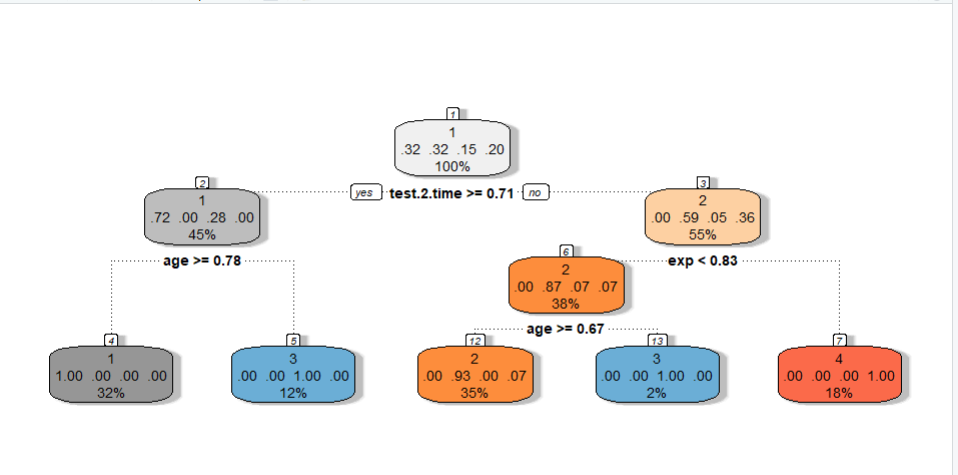
> Train <- subset(data2, split == TRUE)

> Test <- subset(data2, split == FALSE)

> data2Tree <- rpart(groups ~ age+test.1.time+test.2.time+exp+education+cert+exp,

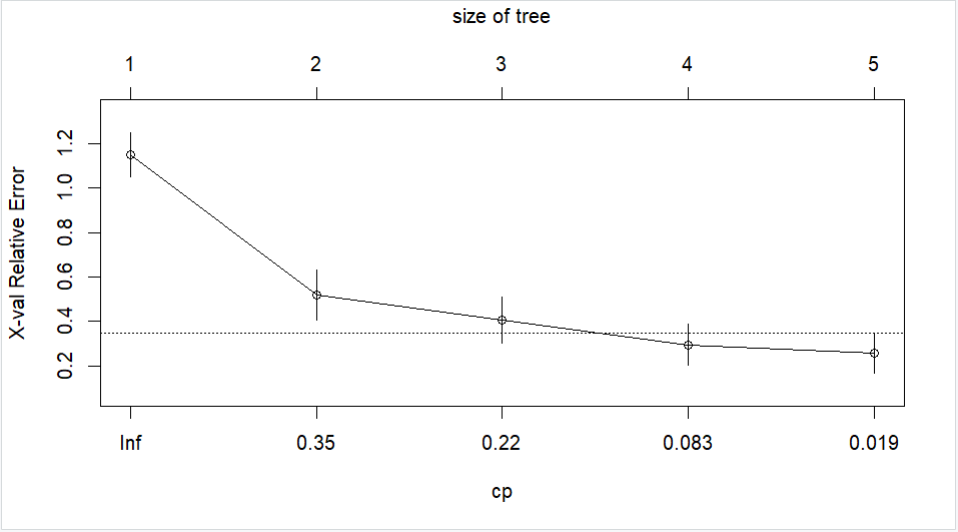
+ data = Train, method = "class", control=rpart.control(minbucket = 1))

> fancyRpartPlot(data2Tree, palettes=c("Greys", "Oranges","Blues", "Reds"))

**

*Рис. 8 Дерево классификации с выделенной тестовой выборкой в R*

Вывод: при построении дерева классификации с выделенной тестовой выборкой в R мы получили хорошо сбалансированное дерево со сложностью ниже, чем в вариантах построения в Statistica.

**

*Рис. 9 График изменения относительной точности дерева в зависимости от сложности модели и размера дерева*

Дерево в R с использованием кросс-валидации:

> library(e1071)

> number=30

> fitControl <- trainControl(method="cv", number=30)

> cartGrid <- expand.grid(.cp=(1:200)\*0.001)

> train(groups ~ age+test.1.time+test.2.time+exp+education+cert, data=data2, method="rpart", trControl=fitControl, tuneGrid=cartGrid)

CART

50 samples

6 predictor

4 classes: '1', '2', '3', '4'

No pre-processing

Resampling: Cross-Validated (30 fold)

Summary of sample sizes: 47, 48, 49, 49, 48, 49, ...

Resampling results across tuning parameters:

cp Accuracy Kappa

0.001 0.8601190 0.7447090

0.002 0.8601190 0.7447090

0.003 0.8601190 0.7447090

0.004 0.8601190 0.7447090

0.005 0.8601190 0.7447090

0.006 0.8601190 0.7447090

…

0.151 0.8601190 0.7447090

…

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.151.

> dataTree3 <- rpart(groups ~ age+test.1.time+test.2.time+exp+education+cert, data = data2, method = "class", control=rpart.control(cp = 0.151))

> fancyRpartPlot(dataTree3, palettes=c("Greys", "Oranges","Blues", "Reds"))

> PredictCART3 <- predict(dataTree3, newdata = data2, type="class")

> table(data2$groups, PredictCART3)

PredictCART3

1 2 3 4

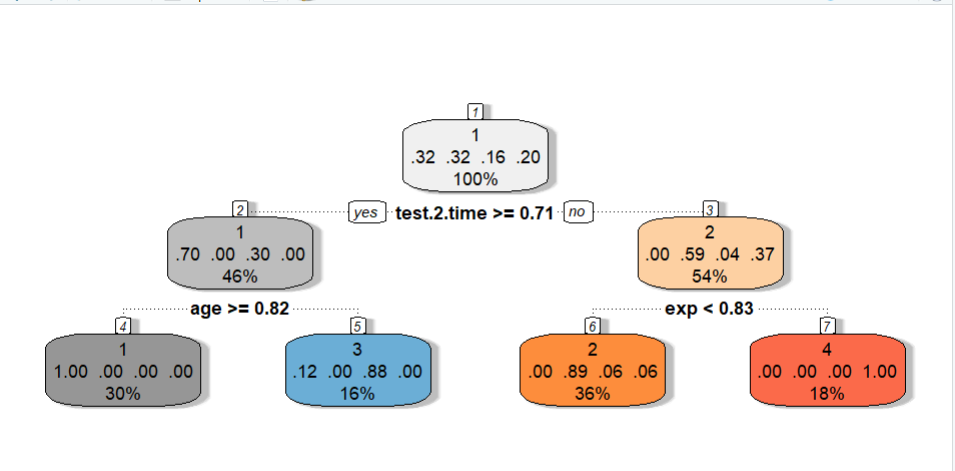
1 15 0 1 0

2 0 16 0 0

3 0 1 7 0

4 0 1 0 9

> plotcp(dataTree3)

**

*Рис. 10 Дерево классификации, построенное с использованием кросс-валидации*

Вывод: использование кросс-валидации при построении в R дало сбалансированное дерево с наименьшей сложностью. Показатели точности возросли по сравнению с вариантом 4.

Анализ вариантов построения деревьев классификации:

*Таблица 1. Показатели качества деревьев решений (Statistica)*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **номер вар. дерева** | **признаки** | **правило останова** | **точность дерева** | | | | **сложность дерева** | | | |
| **Resub.cost\*** | **CV-cost+s.d.** | **Global CV- cost +s.d.** | **Test-cost+s.d.** | **число листьев** | **число вершин** | **максимальн. длина пути** | **длина внеш- него пути** |
| **1** | Все колич. и кач. признаки | Прямая остановка по методу FACT | 0,306548 | 0,417 | - | - | 7 | 6 | 4 | 21 |
| **2** | Все колич. и кач. признаки | Отсечение по ошибке классификации | 0,306548 | 0,435 | 0,47578 | - | 4 | 3 | 3 | 9 |
| **3** | Все колич. и кач. Признаки + тестовая выборка | Прямая остановка по методу FACT | 0,348485 | 0,532 | 0,55199 | 0,298 | 7 | 6 | 4 | 21 |

*Таблица 2. Показатели качества деревьев решений (R)*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **номер**  **вар. дерева** | **точность дерева** | | | | | **сложность дерева** | | | |
| **Цена обуче- ния** | **accuracy** | **precision** | **AUC** | **F-мера** | **число листьев** | **число вершин** | **максимальн. длина пути** | **длина внешнего пути** |
| **4** | 0.1 | 0.9 | 0.9375 | 0.917 | 0.8809 | 5 | 3 | 3 | 12 |
| **5** | 0.06 | 0.94 | 0.9409 | 0.975 | 0.9328 | 4 | 2 | 2 | 8 |

По итогам построения деревьев, лучший вариант построения – 5, т.к. у него наименьшая сложность и наибольшая точность.

При смене правил останова возрастает цена кросс-проверки, но уменьшается сложность построенного дерева.

1. **Вывод**

В ходе выполнения работы были изучены алгоритмы и методы деревьев классификации, исследована их эффективность при решении прикладной задачи. Получены навыки работы с модулями реализации методов деревьев классификации в Statistica и языке R.

* Наилучшим вариантом классификации получилось дерево, построенное с использованием кросс-валидации.
* Полученное дерево имеет сравнительно низкую сложность и наименьшее число ошибок.