Министерство образования и науки Российской Федерации
Санкт-Петербургский государственный технический университет
Институт прикладной математики и механики
Кафедра «Телематика»

ЛАБОРАТОРНАЯ РАБОТА

ПО ТЕМЕ

«Деревья решений»

по направлению 02.04.01.02 «Организация и управление суперкомпьютерными системами»

Выполнил:
Студент гр. 13643.1 Титов А.И.
Проверил: Уткин Л.В.

Санкт-Петербург
2019

Оглавление

Постановка задачи

Требуется выполнить следующие задачи:

1. Загрузить набор данных Glass из пакета “mlbench”. Построить дерево классификации для модели, задаваемой следующей формулой: $Type \sim .$, дайте интерпретацию полученным результатам. При рисовании дерева использовать параметр cex=0.7 для уменьшения размера текста на рисунке. Выявить, является ли построенное дерево избыточным. Выполнить все операции оптимизации дерева. Определить, к какому типу стекла относится экземпляр с характеристиками:

   RI =1.516 Na =11.7 Mg =1.01 Al =1.19 Si =72.59 K=0.43 Ca =11.44 Ba =0.02 Fe =0.1

2. Загрузить набор данных spam7 из пакета DAAG. Построить дерево классификации для модели, задаваемой следующей формулой: $yesno \sim .$, дать интерпретацию полученным результатам. Запустить процедуру “cost-complexity prunning” с выбором параметра k по умолчанию, method = ’misclass’, вывести полученную последовательность деревьев. Выявить, какое из полученных деревьев является оптимальным.

3. Загрузить набор данных nsw74psid1 из пакета DAAG. Построить регрессионное дерево для модели, задаваемой следующей формулой: $re78 \sim ..$

4. Загрузить набор данных Lenses Data Set из файла Lenses.txt:

   - 3 класса (последний столбец):
     1 : пациенту следует носить жесткие контактные линзы;
     2 : пациенту следует носить мягкие контактные линзы;
     3 : пациенту не следует носить контактные линзы.
   - Признаки (категориальные):
     Возраст пациента: (1) молодой, (2) предстарческая дальнозоркость, (3) старческая дальнозоркость;
     Состояние зрения: (1) близорукий, (2) дальнозоркий
     Астигматизм: (1) нет, (2) да
     Состояние слезы: (1) сокращенная, (2) нормальная

   Построить дерево решений. Какие линзы надо носить при предстарческой дальнозоркости (2), близорукости (1), при наличии астигматизма (2) и сокращенной слезы (1)?

5. Для построения классификатора используйте заранее сгенерированные обучающие и тестовые выборки, хранящиеся в файлах svmdata4.txt, svmdata4test.txt.

6. Разработать классификатор на основе дерева решений для данных Титаник (Titanic dataset).

1 Набор данных «Glass»

Для обучающего множества было построено дерево решений (рис. 1). При отсутствии ограничений на глубину дерева и 20 терминальных вершинах в качестве максимального количества терминальных вершин было выявлено, что приведенный пример принадлежит второму классу с вероятностью ≈ 0.83.

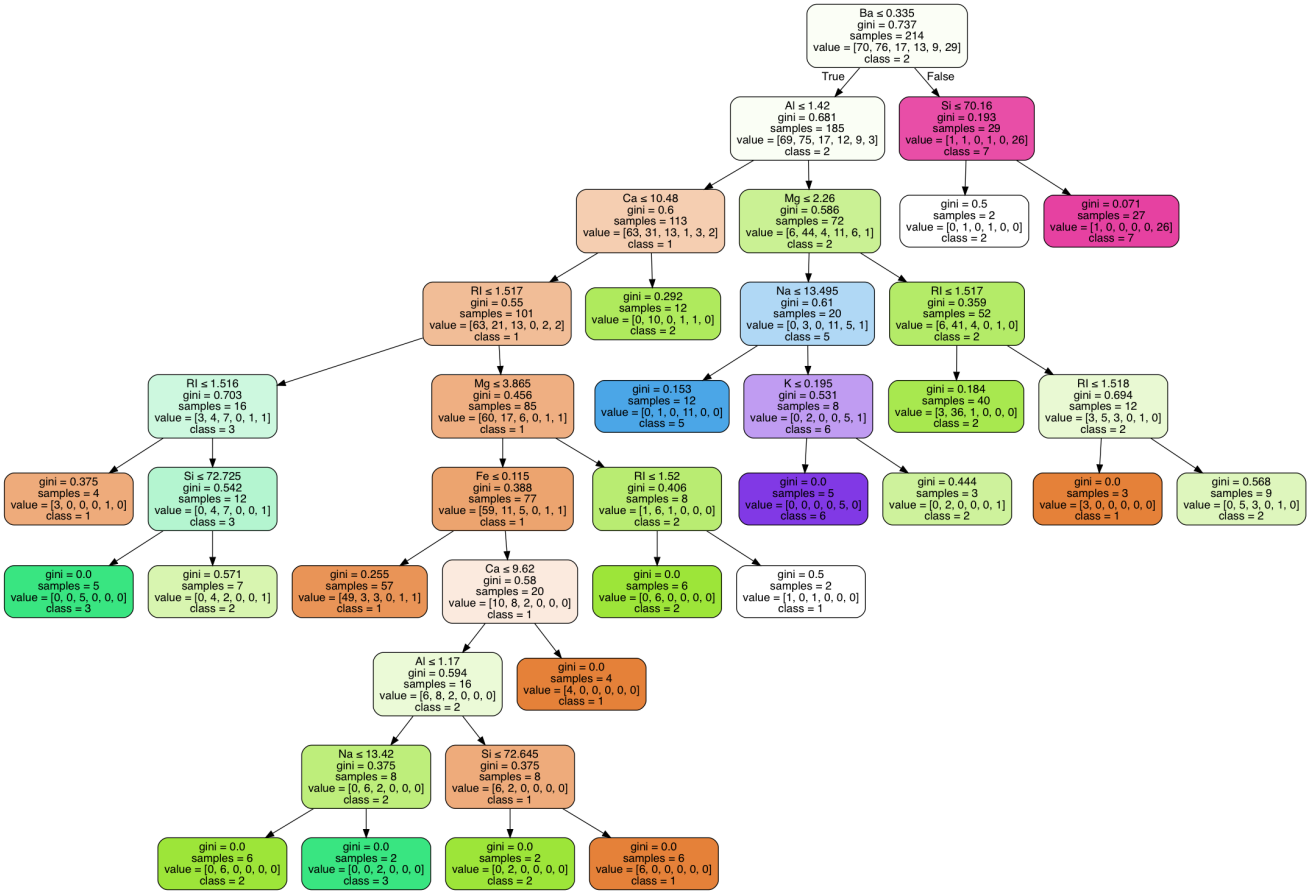

Рис. 1. Дерево решений для примера «Glass»

Однако ограничив глубину дерева до 4ех было достигнуто значение вероятности 1.0. При этом появилась возможность сократить количество терминальных вершин до 17ти (рис. 2).

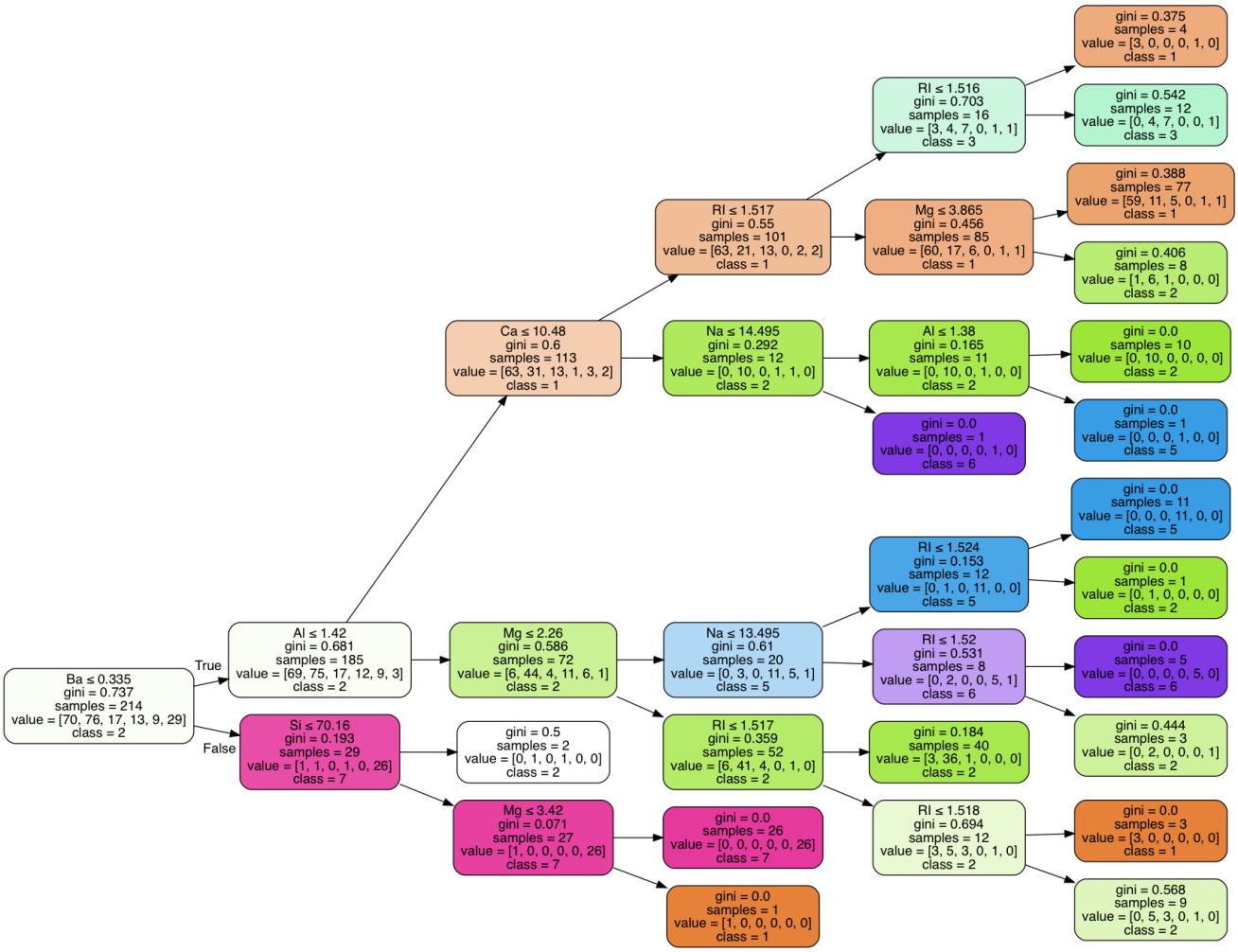

Рис. 2. Усеченное дерево решений для примера «Glass»

2 Набор данных «spam7»

Было построено дерево решений (рис. 3).

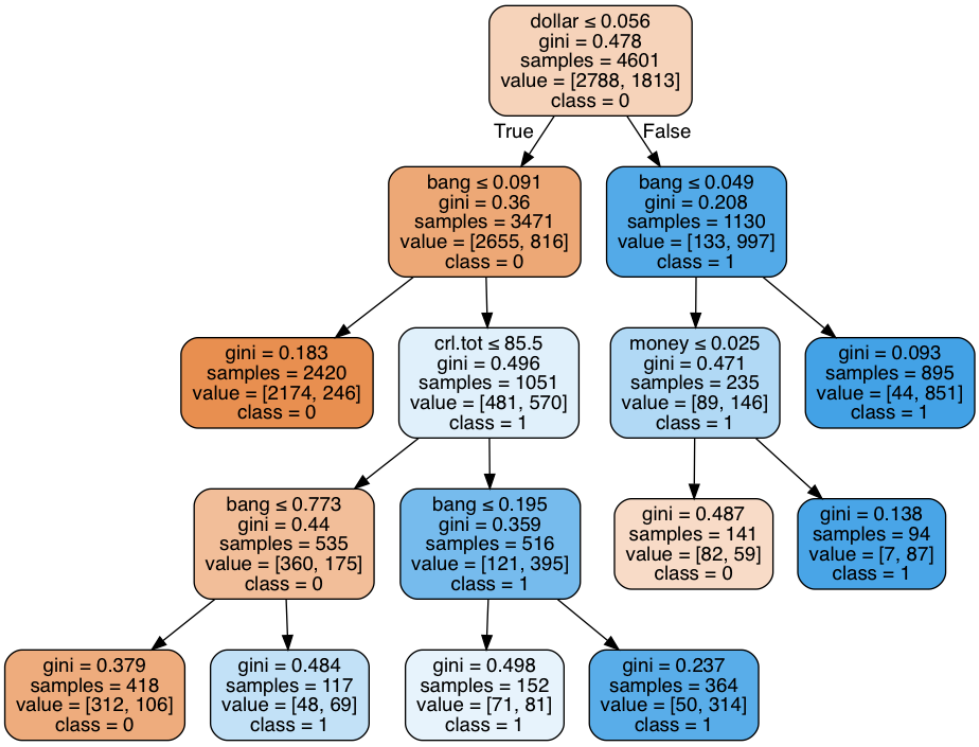

Рис. 3. Дерево решений для примера «spam7»

Для того, чтобы определить наиболее удачное количество терминальных вершин был построен график зависимости ошибочных предсказаний от количества терминальных вершин (рис. 4). Проанализировав график можно сделать вывод, что наилучшим выбором будут значения от 5 до 7. Далее анализируя результаты экспериментов было выбрано значение 6 в качестве оптимального.

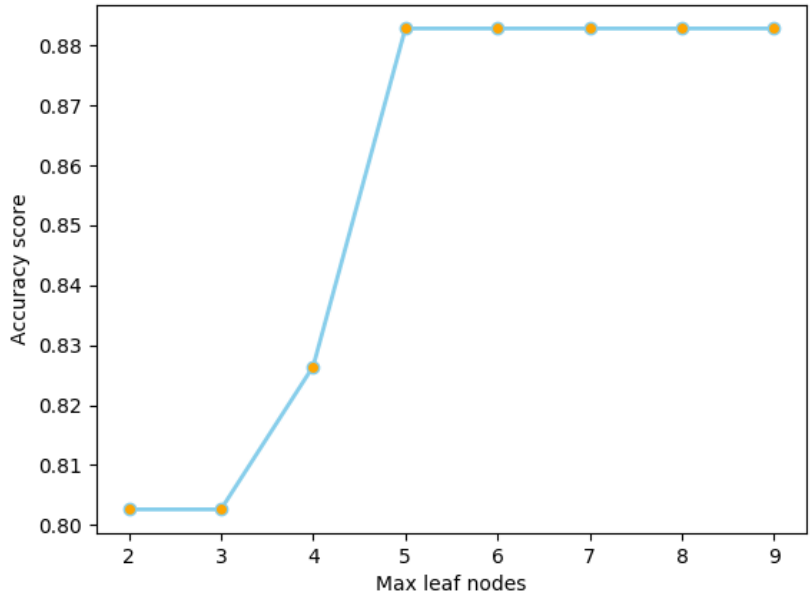

Рис. 4. График зависимости числа ошибок от количества терминальных вершин

Полученное дерево изображено на рисунке (рис. 5).

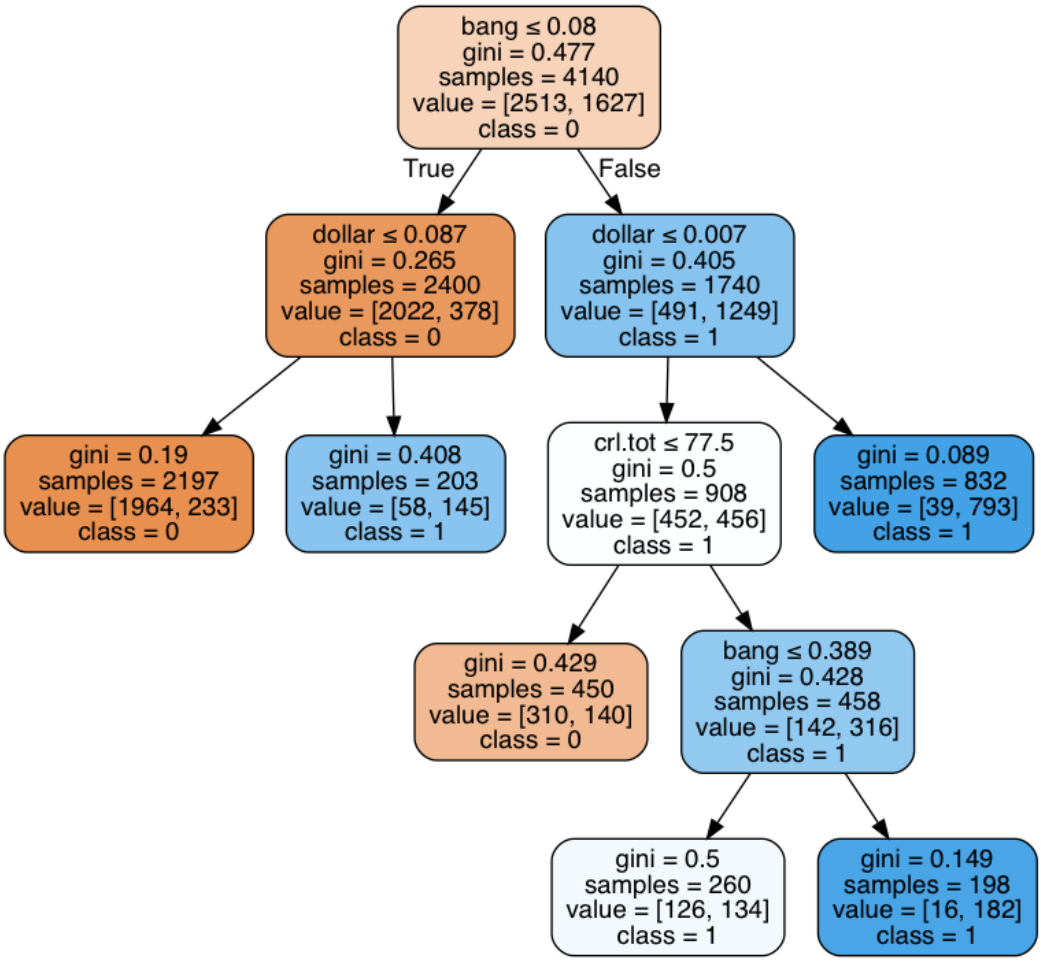

Рис. 5. Измененное дерево решений для примера «spam7»

3 Набор данных «nsw74psid1»

Для приведенного набора было построено регрессионное дерево решений (рис. 6)

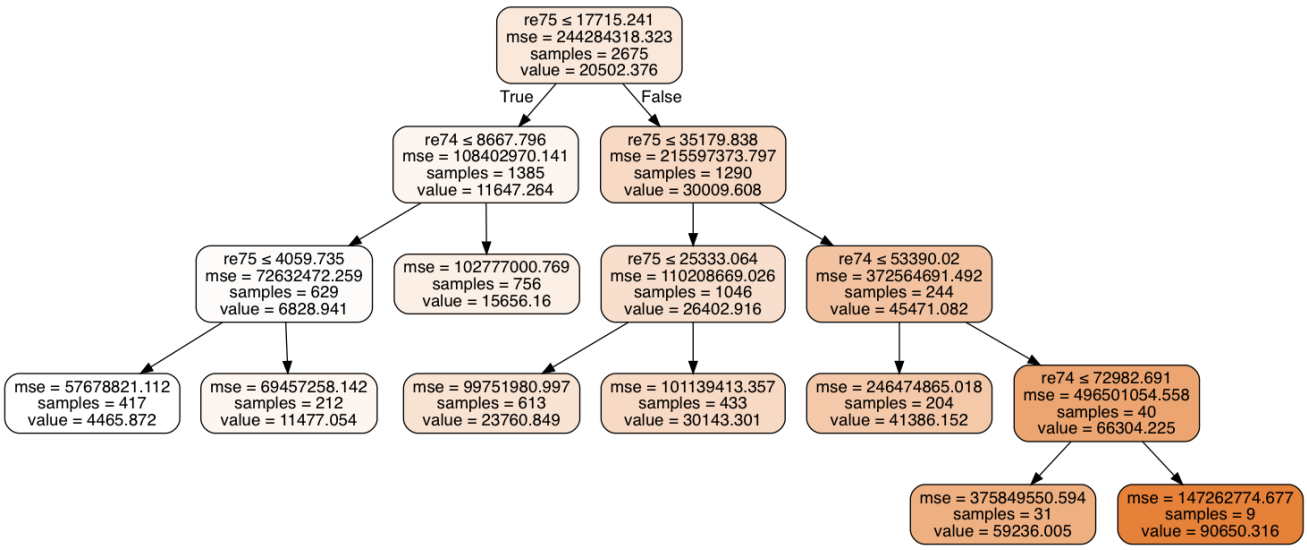

Рис. 6. Дерево решений для примера «nsw74psid1»

4 Набор данных «lenses»

Для набора данных было построено дерево решений (рис. 7).

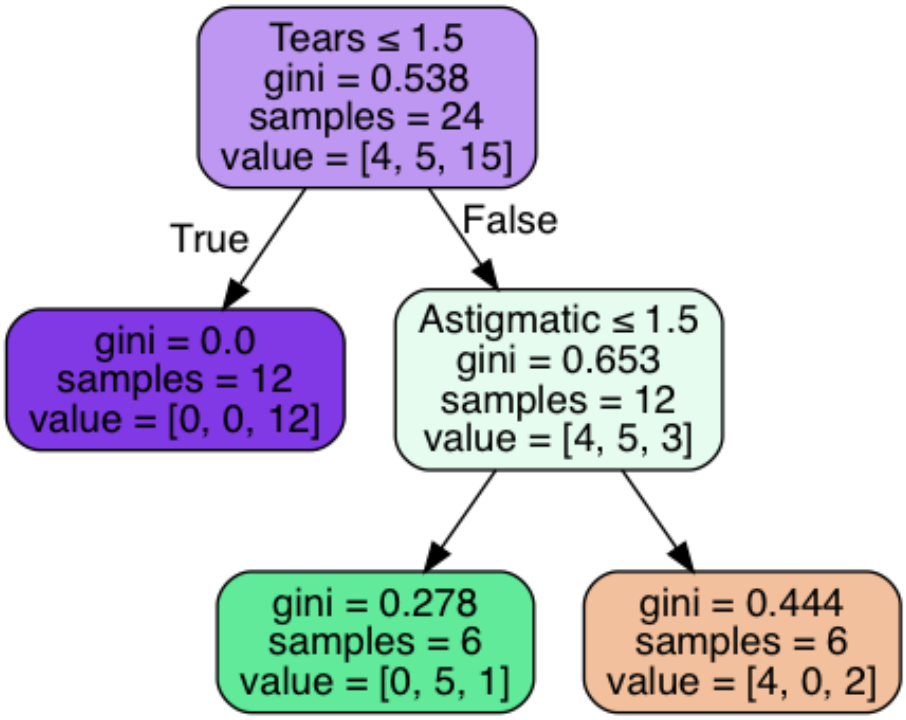

Рис. 7. Дерево решений для примера «lenses»

Также было установлено что при предстарческой дальнозоркости (2), близорукости (1), при наличии астигматизма (2) и сокращенной слезы (1) не следует носить контактные линзы с вероятностью 1.0.

5 Набор данных «svmdata4»

Для набора данных было построено дерево решений (рис. 8).

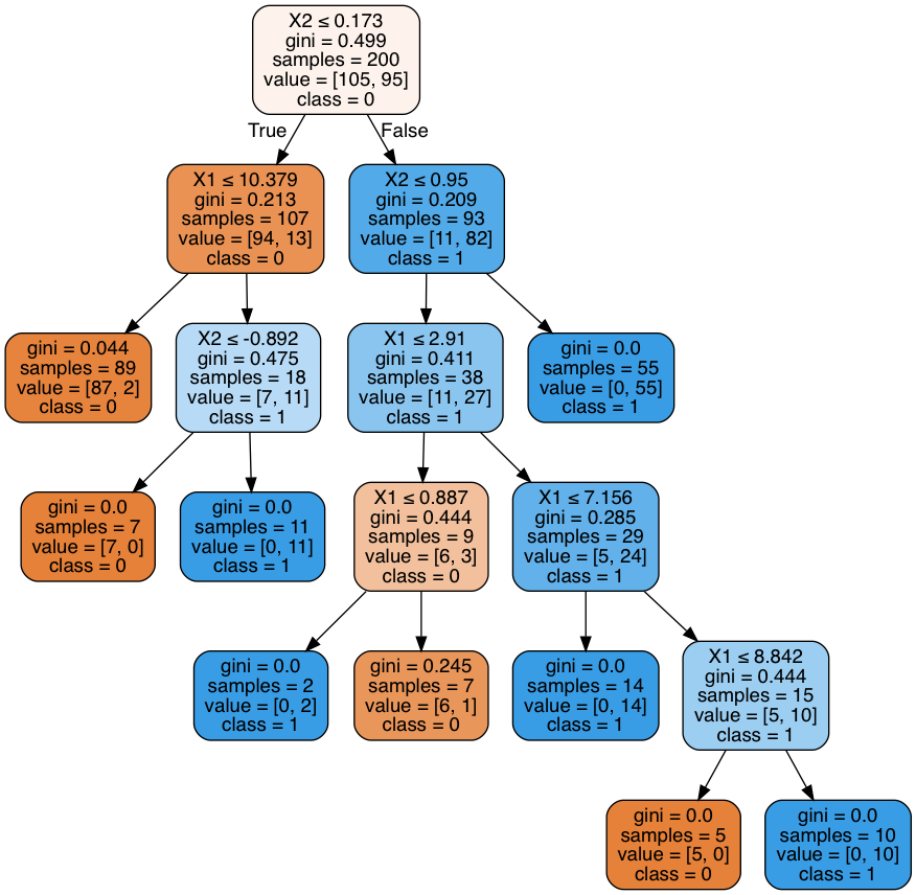

Рис. 8. Дерево решений для примера «svmdata4»

Вероятность ошибки составила 0.1. Ниже представлена таблица отображающая результаты сравнения полученных данных с исходными.

	0	**1**
0	98	3
1	17	82

Таблица 1. Сравнение результатов с исходными данными (svmdata4))

6 Набор данных Titanic

В данном примере обучение производилось по всем возможным колонкам, за исключением колонки «Ticket», а также колонки «SibSp» и «Parch» были заменены на их сумму +

1 (колонка «FamilySize»), колонка «Cabin» была заменена на колонку с бинарными значениями «HasCabin».

Полученное дерево изображено ниже (рис. 9)

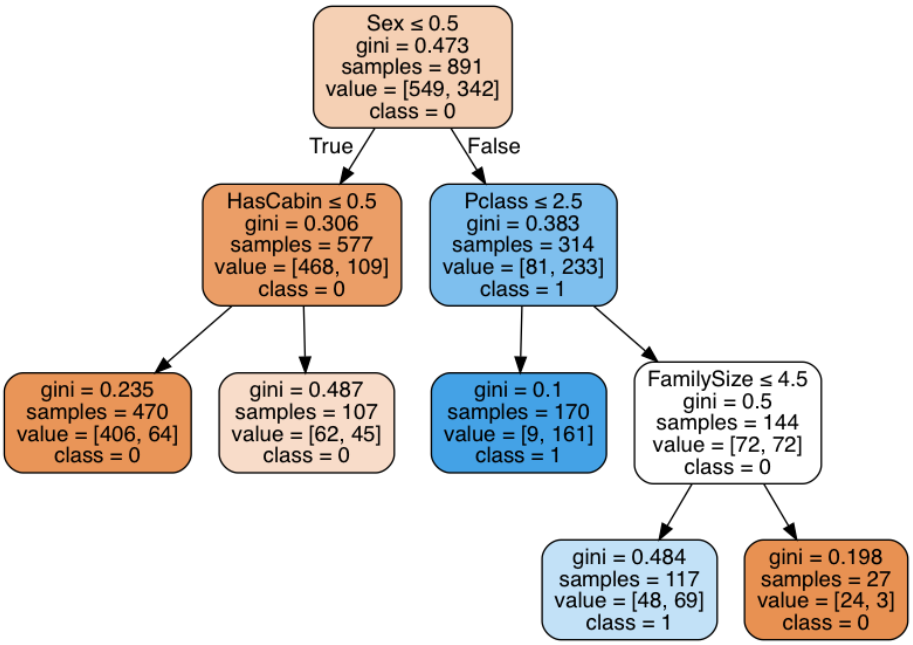

Рис. 9. Дерево решений для примера «Titanic»

Был произведен подсчет величины ошибочной классификации (таблица 9).

	0	**1**
0	266	0
1	5	147

Таблица 2. Сравнение результатов с исходными данными (Titanic dataset))

Таким образом величина ошибочной классификации в примере составила ≈ 0.01.