САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ им. ПЕТРА ВЕЛИКОГО

Институт прикладной математики и механики

Высшая школа прикладной математики и вычислительной физики

Отчёт

по лабораторной работе №3 «Деревья решений» по дисциплине «Системы искусственного интеллекта»

Студентка гр. 3630201/70101 _____ О. В. Саксина

Преподаватель _____ Л. В. Уткин

Санкт-Петербург,
2021г.

Содержание

1 Задание 1

1.1 Постановка задачи

Загрузить набор данных Glass из пакета “mlbench”. Построить дерево классификации для модели, задаваемой следующей формулой: Type ∼ ., дать интерпретацию полученным результатам. 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

1.2 Реализация

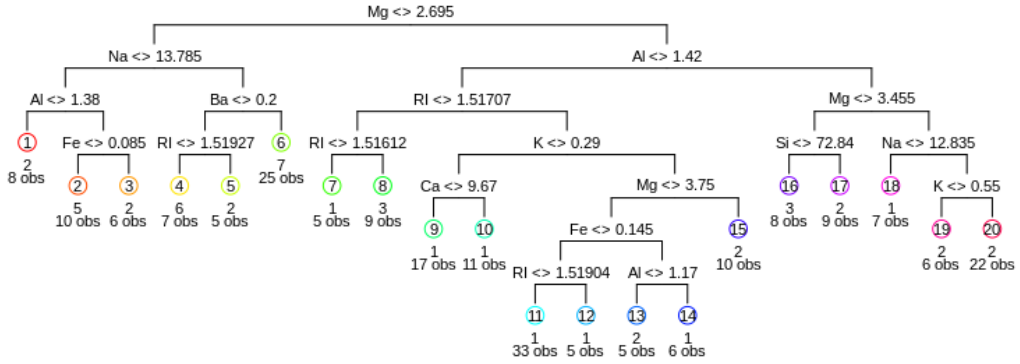

Рис. 1: Дерево решений для набора данных Glass

На Рис. 1 представлено полученное дерево решений. Это дерево является избыточным, т.к. присутствует разделение внутри одного класса в листьях 9 и 10, 11 и 12, 19 и 20. На Рис. 2 представлено дерево решений после оптимизации.

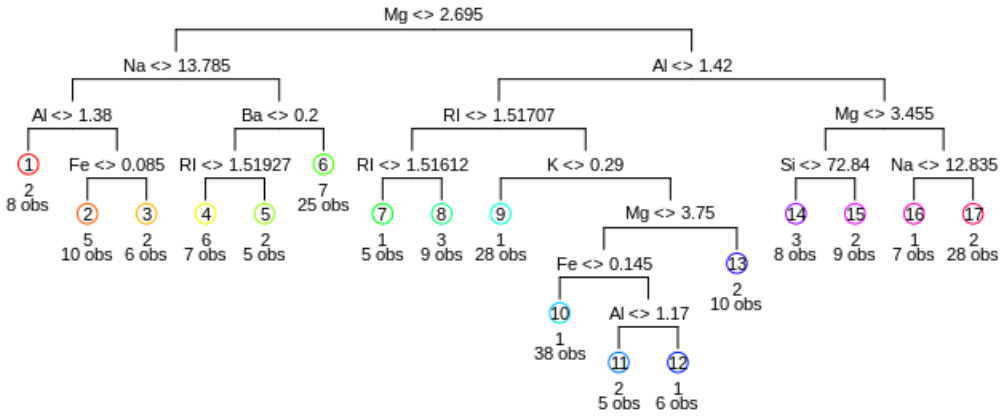

Рис. 2: Оптимизированное дерево решений

Можно заметить, что признак Ca не присутствует в оптимизированном дереве, то есть содержание кальция не помогает классифицировать тип стекла.
По этому дереву видно, что стекло 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 принадлежит к типу 2.

2 Задание 2

2.1 Постановка задачи

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

2.2 Реализация

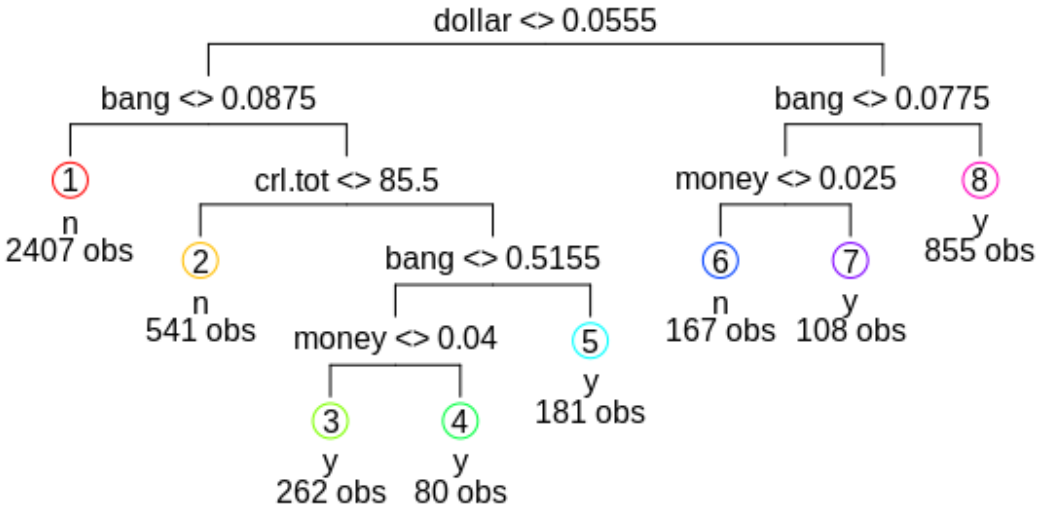

Рис. 3: Дерево решений для набора данных spam7

На Рис. 3 представлено полученное дерево решений. Один и тот же класс в листьх 3 и 4 указывает на то, что дерево избыточно. Оптимизированное дерево представлено на Рис. 3.

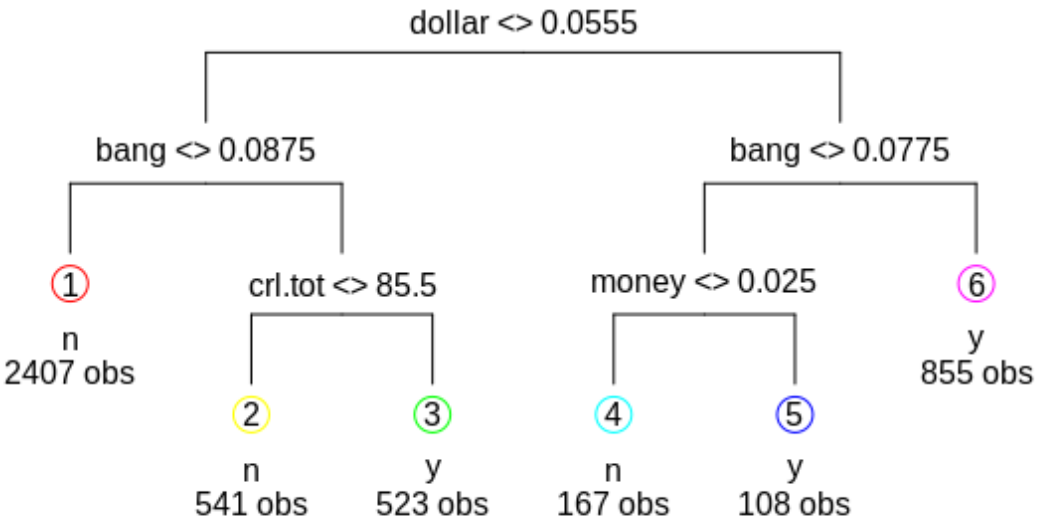

Рис. 4: Оптимизированное дерево решений для набора данных spam7

На классификацию не влияют признаки n000 (количество строк '000') и make (количество слов 'make'). После выполнения процедуры cost-complexity prunning получены деревья, представленные на Рис. 5-7.

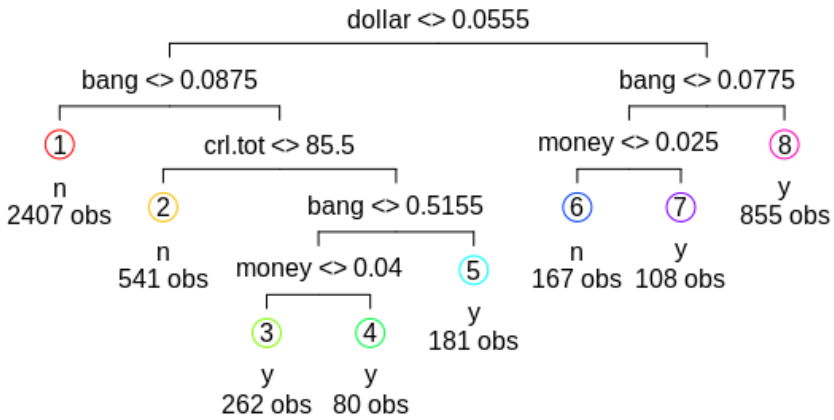

Рис. 5: Оптимизированное дерево решений 1

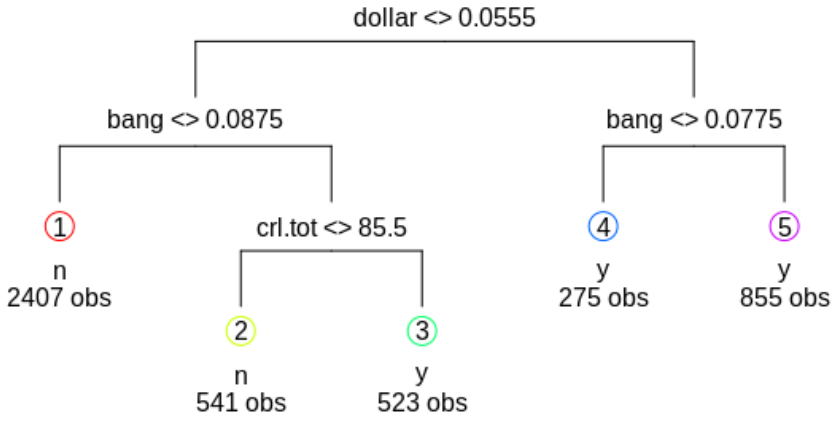

Рис. 6: Оптимизированное дерево решений 2

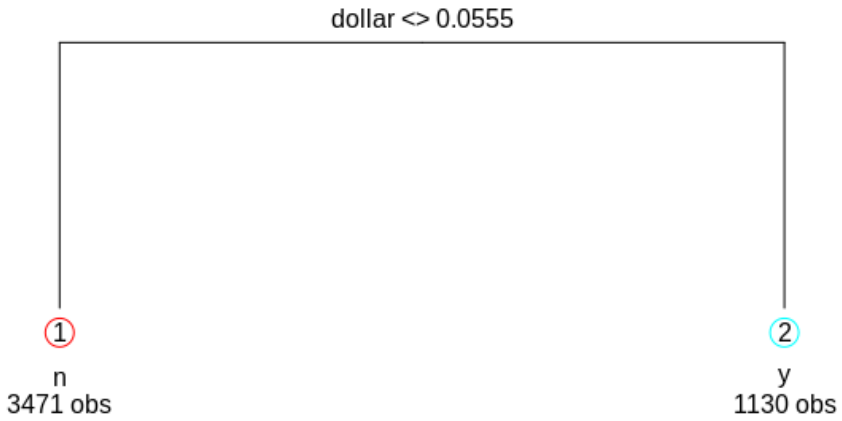

Рис. 7: Оптимизированное дерево решений 3

Дерево 1 совпадает с изначально построенным деревом, дерево 3 классифицирует письма только по одному признаку, что не может быть оптимальным решением в данной задаче, дерево 2 похоже на сокращённое дерево на Рис. 4, но в нём присутствует расщепление по одному классу. На этом основании можно сделать вывод о том, что наиболее оптимальным деревом является дерево на Рис. 4, точность классификации составляет на этой модели 85%.

3 Задание 3

3.1 Постановка задачи

Загрузить набор данных nsw74psid1 из пакета DAAG. Построить регрессионное дерево для модели, задаваемой следующей формулой: re78 ∼.. Построить регрессионную модель и SVM-регрессию для данной формулы. Сравнить качество построенных моделей, выбрать оптимальную модель и объяснить свой выбор.

3.2 Реализация

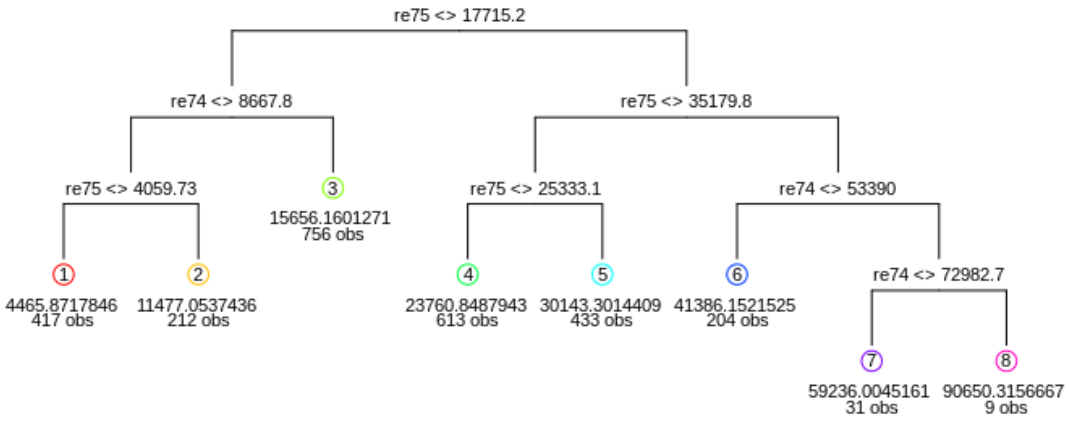

Рис. 8: Дерево решений для набора данных nsw74psid1

На рис. 8 представлено дерево решений для набора данных nsw74psid1. Видно, что признак re78 определяется только по признакам re74 и re75, то есть величина зарплаты в 1978 году у опрошенных не зависела ни от возраста, ни от уровня образования, ни от семейного положения, ни от цвета кожи, а только от величины зарплаты в 1974-1975 годах.
Была построена SVM-регрессия для тех же данных. Результаты распределения данных представлены на рис. 9 и 10 (красным обозначены данные nsw74psid1, а зелёным и жёлтым – предсказания моделей) Средняя ошибка классификации SVM-регрессии равна 6561. Средняя ошибка классификации регрессионного дерева решения равна 7349. То есть SVM-регрессия является более оптимальной.

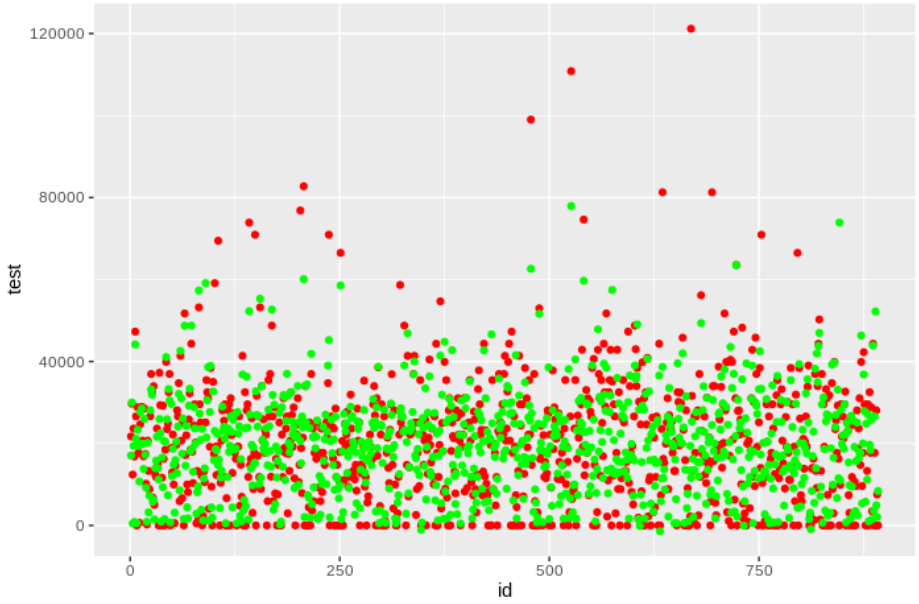

Рис. 9: Данные, полученные после обучения регрессионной SVM модели

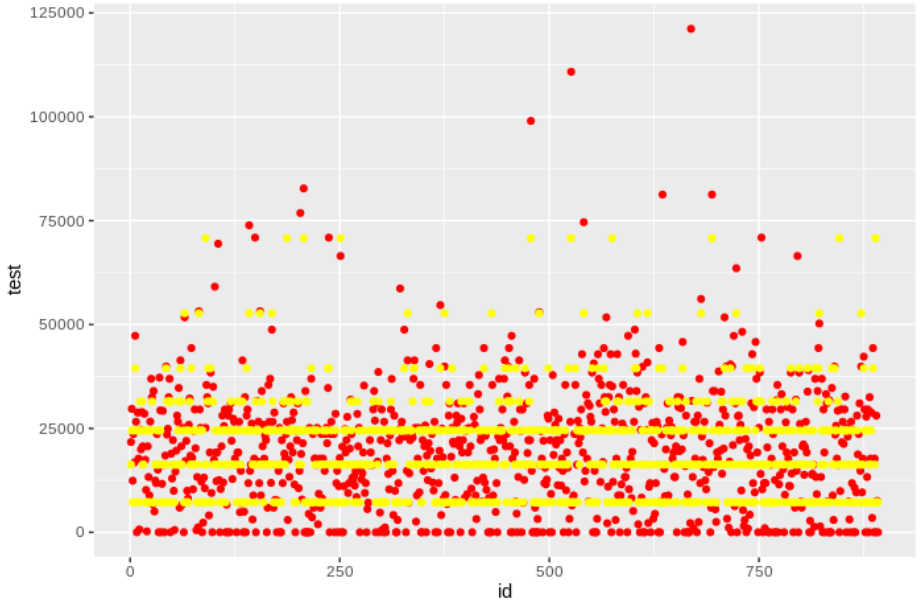

Рис. 10: Данные, полученные после обучения модели регрессионного дерева решения

4 Задание 4

4.1 Постановка задачи

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):
1. возраст пациента: (1) молодой, (2) предстарческая дальнозоркость, (3) старческая дальнозоркость
2. состояние зрения: (1) близорукий, (2) дальнозоркий
3. астигматизм: (1) нет, (2) да
4. состояние слезы: (1) сокращенная, (2) нормальная
Построить дерево решений. Какие линзы надо носить при предстарческой дальнозоркости, близорукости, при наличии астигматизма и сокращенной слезы?

4.2 Реализация

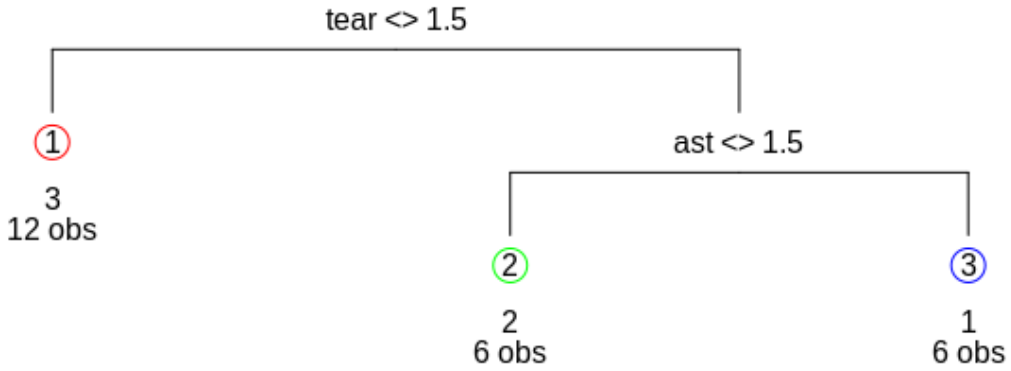

Рис. 11: Дерево решений для набора данных Lenses Data Set

На рис. 11 представлено дерево решений для набора данных Lenses Data Set. На классификацию вляют только признаки состояние слезы и астигматизм. В соответсвии с полученным классификатором при предстарческой дальнозоркости, близорукости, при наличии астигматизма и сокращенной слезы не следует носить контактные линзы.

5 Задание 5

5.1 Постановка задачи

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

5.2 Реализация

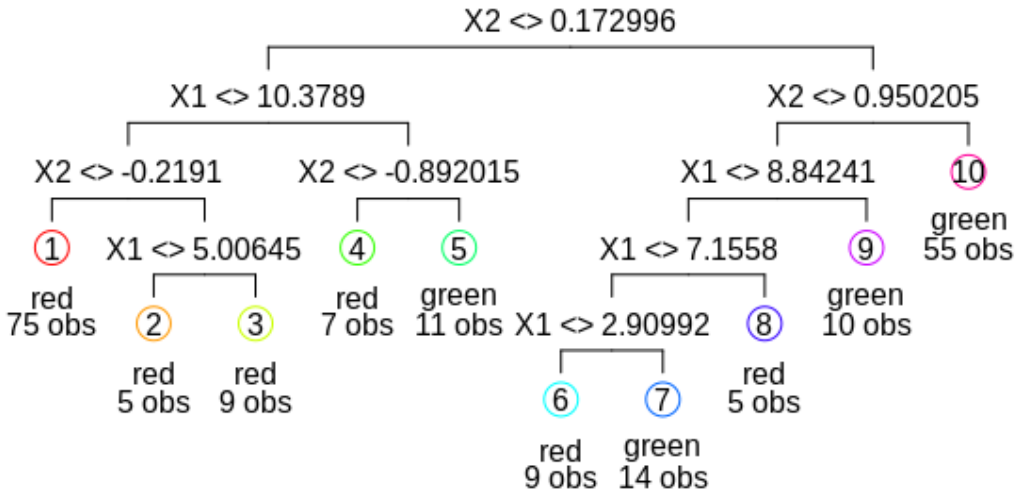

Рис. 12: Дерево решений для набора данных svmdata4

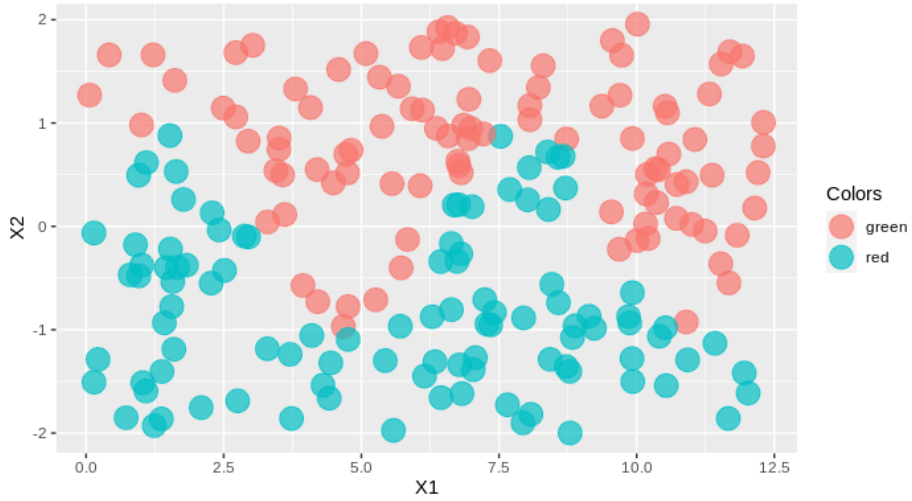

Рис. 13: Тестовый набор данных

На рис. 12 представлено дерево решений для набора данных svmdata4. Ошибка классификации обученной модели на тестовых данных (рис. 13) равна 0.155.

6 Задание 6

6.1 Постановка задачи

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

6.2 Реализация

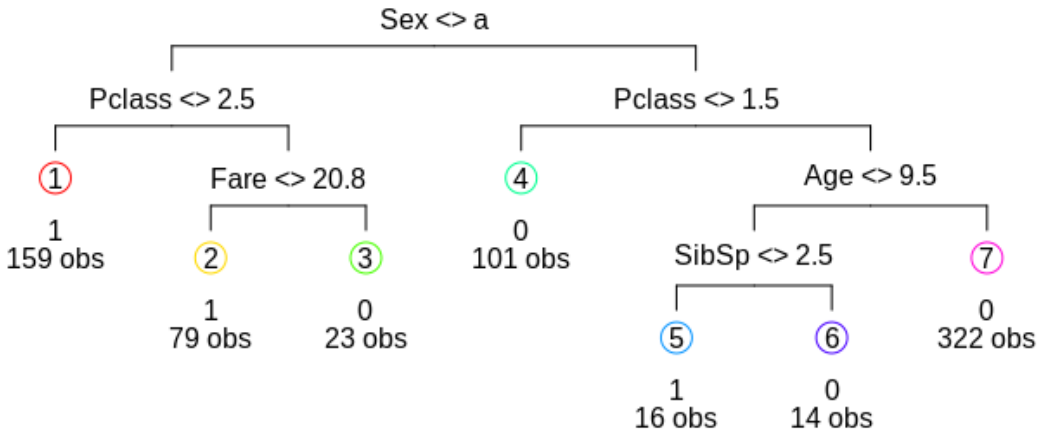

Рис. 14: Дерево решений для набора данных Титаник

Логично предположить, что на вероятность быть спасённым не влияют такие факторы, как идентификатор,имя, номер билета, каюта и порт посадки. Тогда признаки PassengerId, Name, Ticket, Cabin и Embarked можно удалить. На рис. 14 представлено полученное дерево решений. Видно, что решающее значение при принятии решения имеет признак пола, имеют также значение класс пассажира, стоимость билета, возраст и количество родственников. На тестовой выборке точность классификации составила 72%.