Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий и анализа данных |

наименование института

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 5

по дисциплине:

|  |
| --- |
| **Методы анализа данных** |
| **«Построение деревьев решений для задачи классификации»** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил | АСУб-20-2 |  |  |  | Арбакова А.В. |
|  | шифр группы |  | подпись |  | Фамилия И.О. |
| Проверил |  |  |  |  | Бучнев О.С. |
|  | должность |  | подпись |  | Фамилия И.О. |

Иркутск 2022 г

**Цель лабораторной работы:**

1. Выбрать среду программирования для языка Python.
2. Получить набор данных из 200 наблюдений с параметрами, согласно номеру варианта (табл. 1.1). Построить деревья решений и выполнить их анализ. При этом необходимо:

* разделить исходную выборку на две равные части: для обучения и для тестирования;
* обучить дерево решений и вывести его на экран;
* используя функцию рисования решающих поверхностей из лабораторной работы № 3, вывести на экран решающие поверхности и данные для построенного дерева решений для каждой пары признаков;
* оценить точность модели на тестирующей выборке;
* в случае, если точность модели менее 0,9, изменяя значение параметров criterion, max\_depth, min\_samples\_split, min\_samples\_leaf попытаться ее повысить.

1. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

**Язык программирования:** Python

**Номер варианта:** 3

|  |  |  |  |
| --- | --- | --- | --- |
| **№ варианта** | **Количество  кластеров** | **Количество признаков** | **Дисперсия  значений** |
| 3 | 2 | 5 | 5 |

# **Ход выполнения работы**

Одним из важнейших алгоритмов в машинном обучении является алгоритм деревьев решений. Деревья решений относятся к числу самых популярных и мощных инструментов Data Mining, позволяющих эффективно решать задачи классификации и регрессии. В отличие от методов, использующих статистический подход, деревья решений основаны на машинном обучении и в большинстве случаев не требуют предположений о статистическом распределении значений признаков. В основе деревьев решений лежат решающие правила вида «если... то...», которые могут быть сформулированы на естественном языке. Поэтому деревья решений являются наиболее наглядными и легко интерпретируемыми моделями.

В основе работы деревьев решений лежит процесс рекурсивного разбиения исходного множества наблюдений или объектов на подмножества, ассоциированные с классами. Разбиение производится с помощью решающих правил, в которых осуществляется проверка значений атрибутов по заданному условию.

Рекурсивными называются алгоритмы, которые работают в пошаговом режиме, при этом на каждом последующем шаге используются результаты, полученные на предыдущем шаге.

Древовидная модель – это набор правил импликации вида «если...то...иначе», которые просто понять и реализовать. В отличие от регрессии и логистической регрессии, деревья имеют способность обнаруживать скрытые шаблоны, соответствующие сложным взаимодействиям в данных.

Деревья решений – иерархические древовидные структуры, состоящие из решающих правил вида «если... то...» и позволяющие выполнять классификацию объектов. В дереве каждому объекту соответствует единственный узел, дающий решение.

В состав деревьев решений входят два вида объектов – узлы и листья. В узлах содержатся правила, с помощью которых выполняется проверка атрибутов, и множество объектов в данном узле разбивается на подмножества. Листья – это конечные узлы дерева, в которых содержатся подмножества, ассоциированные с классами. Основным отличием узла от листа является то, что в листе не выполняется проверка, разбивающая подмножества, и поэтому нет ветвления.

Начальный узел дерева является входным. Через него проходят все объекты исходного множества. Входной узел называют корневым узлом, и дерево растет сверху вниз. Узлы и листья, подчиненные узлу более высокого иерархического уровня, называются потомками, или дочерними узлами, а тот узел по отношению к ним — предком, или родительским узлом.

Для каждого узла нужно выбрать атрибут, по которому будет выполняться проверка правила. Его принято называть атрибутом разбиения или атрибутом ветвления.

Каждый узел или лист дерева обладает следующими характеристиками:

* количество примеров, попавших в узел (лист);
* доли примеров, относящихся к каждому из классов;
* число классифицированных примеров (для листьев);
* процент примеров, верно классифицированных данным узлом (листом).

Наилучшим разбиением считается то, которое позволяет классифицировать наибольшее число примеров и создавать максимально чистые подмножества, в которых примесь объектов другого класса (то есть не ассоциированного с данным узлом или листом) минимальна.

Эффективность деревьев решений определяется с помощью тестового множества – набора примеров, которые не использовались при построении дерева. Дереву предъявляется набор тестовых примеров и вычисляется, для какого процента примеров класс определен верно. Это позволяет оценить ошибку классификации, а также качество решения задачи классификации или регрессии отдельных ветвей в дереве.

В теории информации существует мера количества информации – энтропия, она отражает степень неупорядоченности системы. При построении деревьев решений под энтропией понимают меру разнообразия классов в узле. Целью разбиения узла в дереве решений является получение дочерних узлов с более однородным классовым составом. В результате разбиения должны образовываться узлы с меньшим разнообразием состояний выходной переменной. Следовательно, энтропия падает, а количество внутренней информации в узле растет. Уменьшение энтропии эквивалентно приросту информации.

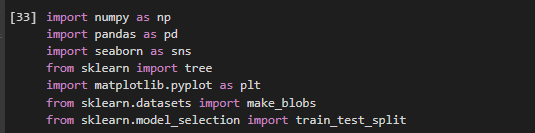
Одним из наиболее популярных алгоритмов построения деревьев решений является алгоритм С4.5. Алгоритм начинает работу со всеми обучающими примерами в корневом узле дерева. Для разделения множества примеров корневого узла выбирается один из атрибутов, и для каждого значения, принимаемого этим атрибутом, строится ветвь и создается дочерний узел. Затем все примеры распределяются по дочерним узлам в соответствии со значением атрибута.

Для реализации алгоритма С4.5 необходимо выполнить – подготовить данные, вычислить энтропию исходного разбиения, сформировать возможные разбиения, вычислить энтропию каждого из возможных разбиений, определить разбиение, соответствующее наибольшему приросту информации, и выполнить разбиение и рекурсивно продолжить итерации (если энтропия какого-то разбиения равна нулю, прекратить его дальнейшее разбиение).

CART (Classification and regression trees) был предложен в 1984 году. Деревья решений, построенные с помощью алгоритма САКТ, являются бинарными, то есть содержат только два потомка в каждом узле.

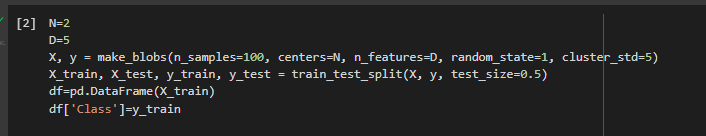
Для реализации алгоритма CART необходимо выполнить – подготовить данные, все признаки необходимо сделать бинарными, сформировать возможные разбиения, вычислить значение показателя для каждого возможного разбиения, выполнить разбиение, соответствующее максимальному значению показателя и рекурсивно продолжать итерации до тех пор, пока значение показателя больше 1.

Подключим необходимые библиотеки и модули:



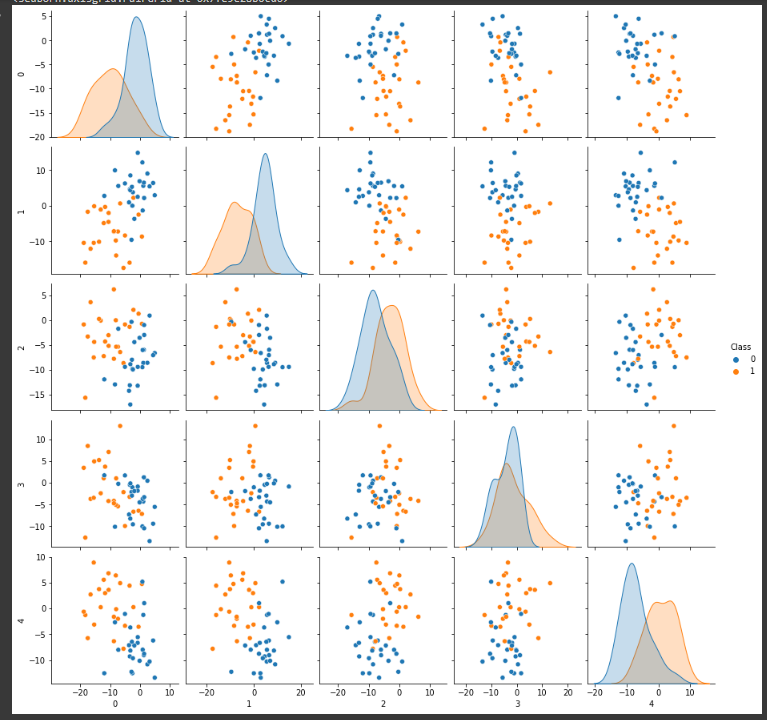
Изображение 1 – Подключаемые библиотеки и модули

Получим исходные данные для числа классов:



Изображение 2 – Получение исходных данных для числа классов

Построим диаграммы рассеяния:



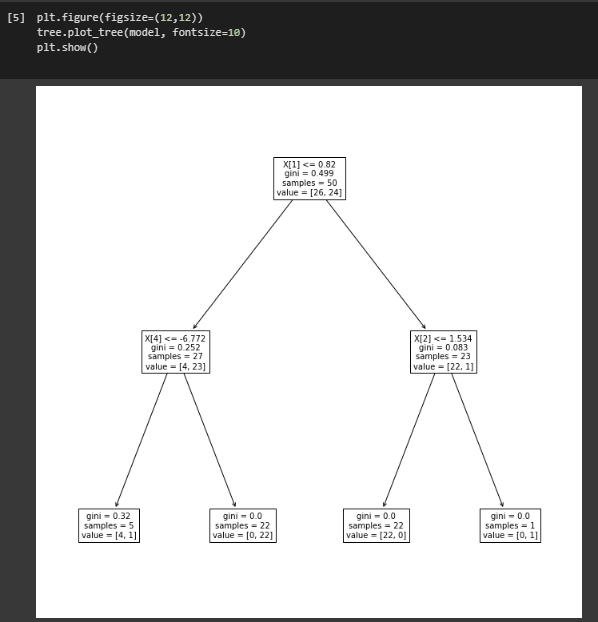
Изображение 3 – Диаграммы рассеяния

Обучим дерево решений:



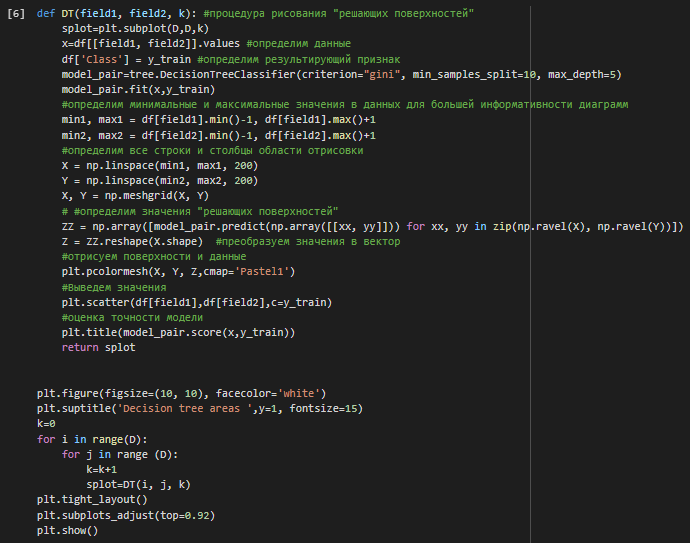
Изображение 4 – Обучение дерева решений

Выведем дерево решений:



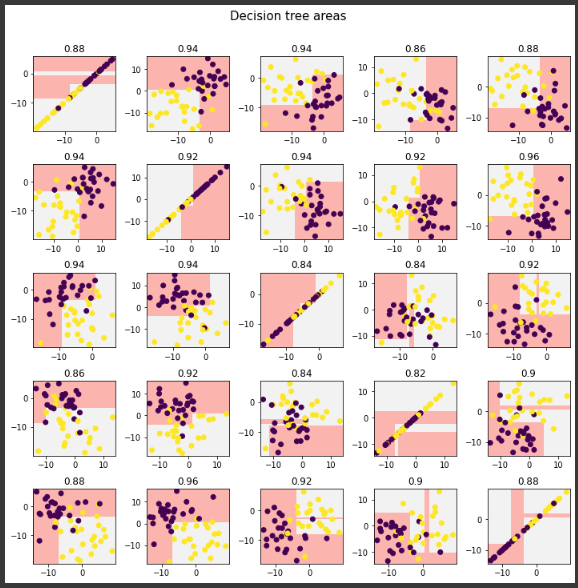
Изображение 5 – Дерево решений

Используем функцию отрисовки решающих поверхностей, построим и выведем на экран деревья решений и решающие поверхности для каждой пары признаков:



Изображение 6 – Код отрисовки решающих поверхностей, постройки деревьев решений и решающих поверхностей для каждой пары признаков

Используем функцию отрисовки решающих поверхностей, построим и выведем на экран деревья решений и решающие поверхности для каждой пары признаков:



Изображение 7 – Отрисовки решающих поверхностей, постройки деревьев решений и решающих поверхностей для каждой пары признаков

Оценим точность модели на тестирующей выборке:



Изображение 8 – Оценка точности модели на тестирующей выборке

# **Заключение**

В ходе выполнения лабораторной работы по теме «Построение деревьев решений для задачи классификации» была выполнена цель работы по построению деревьев решений для задачи классификации.

Было изучены такие темы как: назначение деревьев решений, основное отличие алгоритмов построения деревьев решений, алгоритм C4.5, алгоритм CART и способы борьбы с переобучением модели.

Знания были закреплены во время использования среды разработки Google Colab и использования языка программирования Python, в котором было выполнено задание лабораторной работы.

# **Список литературы**

1. Бучнев О. С. Методы анализа данных: лабораторный практикум – Иркутск, 2022. – 115 с
2. <https://colab.research.google.com>
3. С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью языка R, Хайдельберг – Лондон – Тольятти, 2014 г. – 401 с