Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 4.

ЛОГИЧЕСКИЕ МЕТОДЫ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

**1.Классификации**

**Цели и задачи** Цель лабораторной работы: изучение принципов построения информационных систем с использованием логических методов классификации. Основные задачи: – освоение технологии внедрения алгоритмов на основе решающих списков в приложения; – освоение технологии внедрения алгоритмов на основе решающих деревьев в приложения; – изучение параметров логической классификации; – освоение модификаций логических методов классификации.

1. **Теоретическое обоснование**

Перед выполнением лабораторной работы необходимо ознакомиться с теорией построения логических классификаторов, используя следующие источники: [1-5].

1. **Методика и порядок выполнения работы**
   1. **Учебная задача** В рамках учебной задачи необходимо произвести построение классификатора на основе логического дерева. В качестве набора данных используется набор данных об ирисах Фишера.

1. Подключим библиотеки, которые потребуются для загрузки и первичного анализа данных (рис. 4.1).

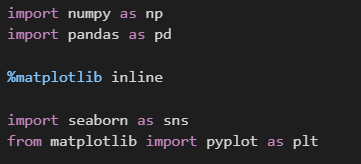


Рисунок 4.1 – Использование pandas для загрузки набора данных

1. Для построения дерева классификации воспользуемся специальным классом sklearn.tree.DecisionTreeClassifier. Оценими точность модели методом hold-out (рис. 4.2). Следует обратить внимание, что если в методе ближайших соседей производилась оптимизация по одному параметру K – количеству ближайших соседей, то при создании модели DecisionTreeClassifier необходимо указать два параметра: максимальную глубину дерева (max\_depth) и количество признаков разделения дерева (max\_features).

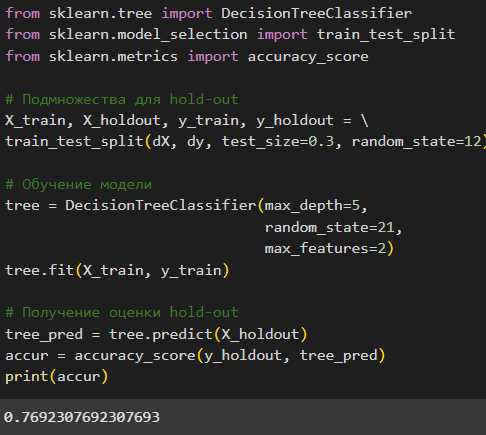
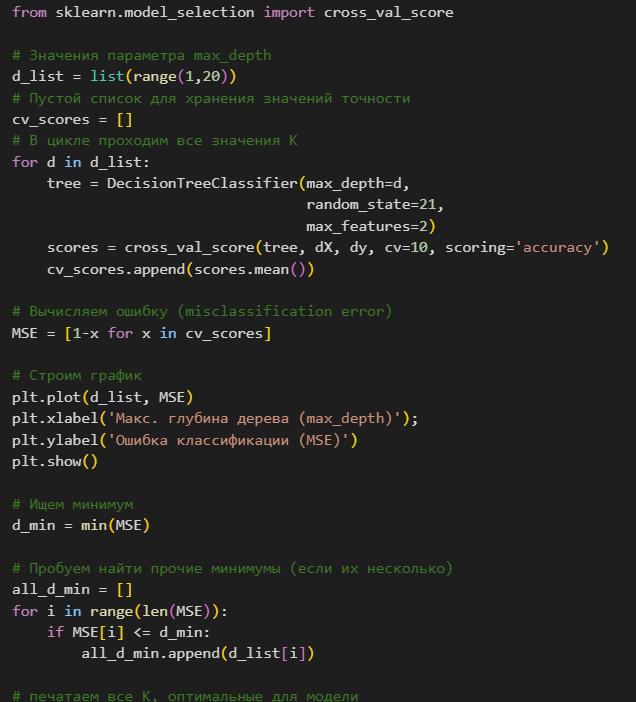


Рисунок 4.2 – Обучение модели классификации и оценка ее точности методом hold-out

1. Произведем оценку точности модели по методу cross validation (рис. 4.3), а также сделаем выводы об оптимальном значении параметра max\_depth. 45

Рисунок 4.3 – Оценка точности модели методом cross validation и нахождение оптимального значения параметра max\_depth

В результате работы данного кода будет получен вывод (рис. 4.4).

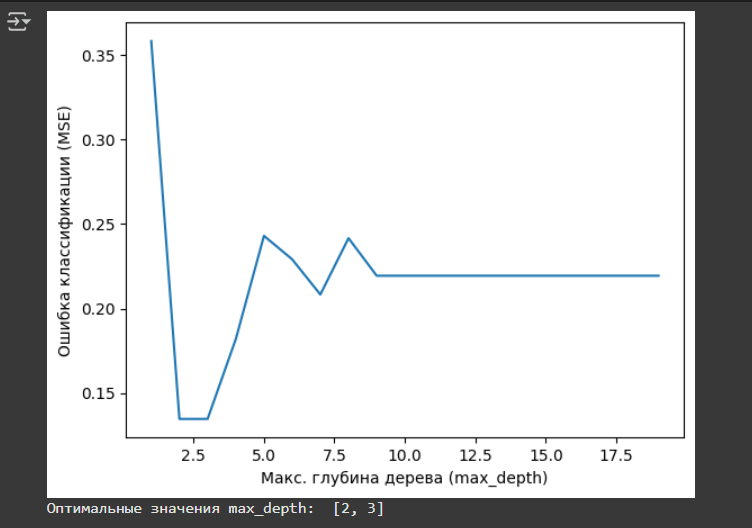
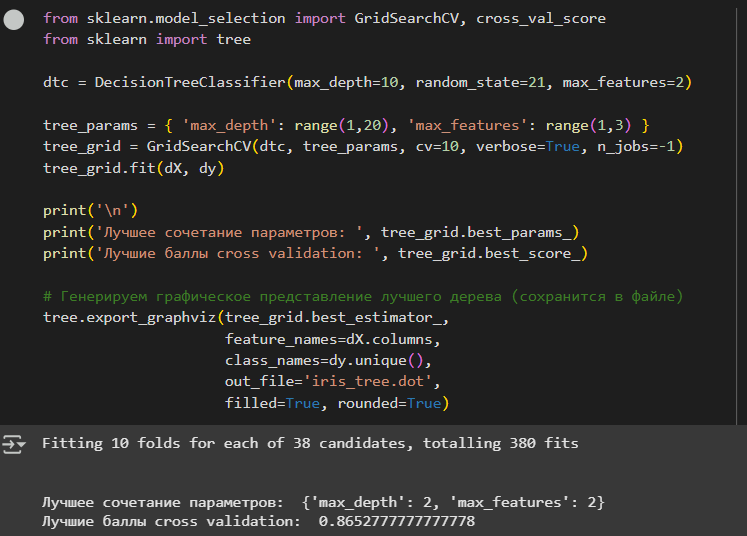


Рисунок 4.4 – Вывод зависимости значения MSE от параметра max\_depth

1. Оптимальное значение параметра max\_depth модели получено, но в модели присутствует еще один параметр max\_features, который был установлен в значение 2 (не изменялся и не оптимизировался). Для проведения cross validation по всем параметрам воспользуемся классом GridSearchCV пакета sklearn.model\_selection (рис. 4.5).



46 Рисунок 4.5 – Находждение оптимальных параметров модели логической классификации

Поясните вывод данного фрагмента. Поясните значение таких величин как fold, candidate, fit. Какие значения принимают данные величины в данном коде и почему? Следует обратить внимание, что в результате оценки оптимальных параметров, фактически, было построено оптимальное дерево классификации. Доступ к данному дереву производится через поле best\_estimator\_ класса GridSearchCV. В коде (рис. 4.5) производится экспорт полученного дерева в формат .dot. Полученный формат можно преобразовать в .png через сервис сайта http://webgraphviz.com. Полученное дерево представлено на рис. 4.6.

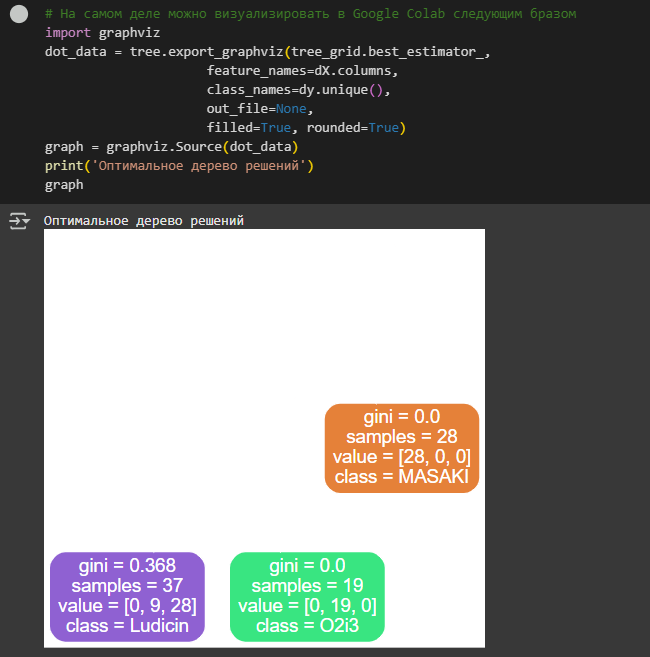


Рисунок 4.6 – Находждение оптимальных параметров модели логической классификации

Поясните значения переменных в узлах полученного дерева: gini, samples, value. 5. Оптимальные параметры определены, можно обучить модель и использовать ее для классификации (рис. 4.7).

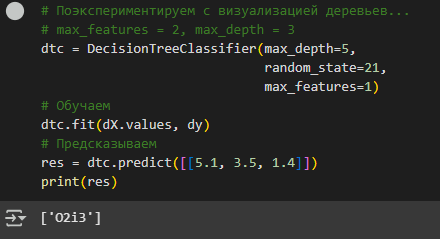


Рисунок 4.7 – Использование модели логической классификации

1. В заключении построим еще одну визуализацию процесса логической классификации – покажем решающие границы модели классификации (рис. 4.8).

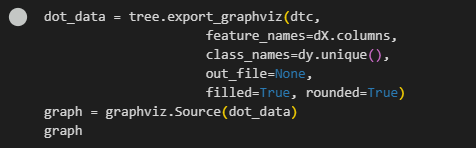


Рисунок 4.8 – Построение решающих границ

Вывод данного кода представлен на рис. 4.9.

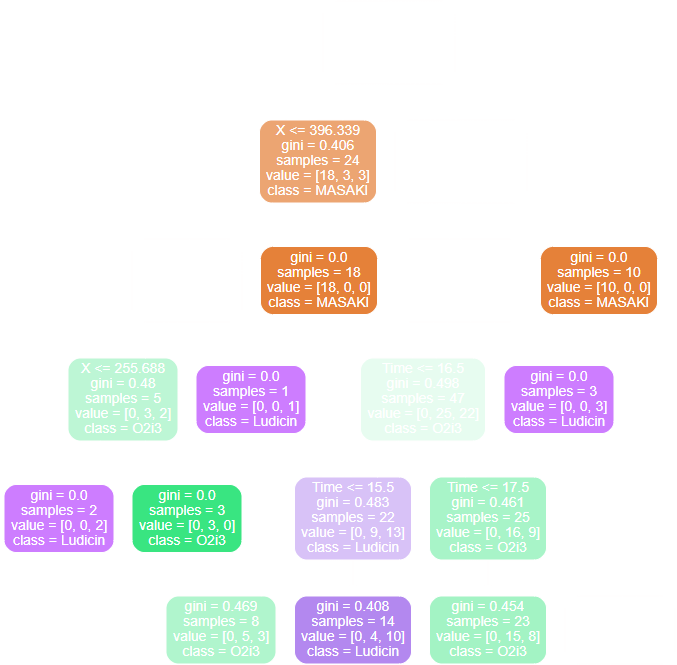


Рисунок 4.9 – Решающие границы логической модели классификации

1. **Контрольные вопросы**

1)Принцип построения дерева решений:

Дерево решений строится рекурсивно, начиная с корня. На каждом шаге выбирается лучший признак для разделения данных на подмножества. Лучший признак определяется с помощью критерия информативности (например, прирост информации или индекс Джини).

Корень: В корне находятся все данные.

Выбор лучшего признака: Выбирается признак, который лучше всего разделяет данные по целевой переменной.

Разделение: Данные разделяются на подмножества в зависимости от значений выбранного признака.

Рекурсия: Для каждого подмножества процесс повторяется, пока не будет достигнут критерий остановки (например, все объекты в подмножестве принадлежат одному классу или достигнута максимальная глубина дерева).

Листья: Листья содержат предсказания для объектов, попавших в этот лист.

2)Статистическое определение информативности:

В статистике информативность признака определяется его способностью разделять данные по целевой переменной. Чем лучше признак разделяет данные, тем он более информативен. Обычно для оценки информативности используются статистические меры, такие как:

Хи-квадрат: Используется для оценки связи между категориальными признаками.

F-статистика: Используется для оценки различий между группами, сформированными по значениям количественного признака.

Коэффициент корреляции: Используется для оценки линейной связи между количественными признаками.

3)Энтропийное определение информативности:

Энтропия — это мера неопределенности. В контексте деревьев решений энтропия используется для оценки однородности подмножеств данных. Чем меньше энтропия подмножества, тем оно более однородно (то есть объекты в подмножестве в основном принадлежат одному классу).

Энтропия: Мера неопределенности в данных.

Прирост информации: Уменьшение энтропии после разделения данных по признаку. Чем больше прирост информации, тем более информативен признак.

Многоклассовая информативность:

Многоклассовая информативность — это мера информативности признака для задач с несколькими классами. Она используется, когда целевая переменная может принимать более двух значений. Для оценки многоклассовой информативности используются разные критерии, такие как:

Прирост информации (для нескольких классов): Обобщение прироста информации для бинарной классификации на случай нескольких классов.

Индекс Джини (для нескольких классов): Обобщение индекса Джини для бинарной классификации на случай нескольких классов.

4)Многоклассовая информативность применяется для выбора лучших признаков при построении моделей классификации с несколькими классами.

5)Бинаризация количественных признаков:

Бинаризация — это преобразование количественного признака в бинарный (то есть принимающий только два значения: 0 или 1). Алгоритм бинаризации:

Выбор порога: Определяется пороговое значение, которое будет использоваться для разделения значений признака на две группы.

Преобразование: Значения признака, меньшие или равные порогу, заменяются на 0, а значения, большие порога, заменяются на 1.

Бинаризация используется для упрощения моделей и повышения их устойчивости к выбросам. Она также может быть полезна, если признак имеет нелинейную связь с целевой переменной.

6)Порядок поиска закономерностей в форме конъюнкций:

Конъюнкция — это логическое "И" (AND). Поиск закономерностей в форме конъюнкций — это поиск правил вида "Если условие 1 И условие 2 И ... И условие N, то следствие". Порядок поиска:

Выбор признаков: Определяются признаки, которые будут использоваться для построения конъюнкций.

Генерация конъюнкций: Генерируются все возможные конъюнкции из выбранных признаков.

Оценка конъюнкций: Оценивается полезность каждой конъюнкции. Полезность может оцениваться с помощью разных мер, таких как поддержка, достоверность и подъем.

Выбор лучших конъюнкций: Выбираются конъюнкции, которые имеют достаточно высокую полезность.

Поиск закономерностей в форме конъюнкций используется для выявления скрытых связей между признаками и построения правил, которые могут быть использованы для предсказания или объяснения поведения системы.

**Вывод:** изучили принципы построения информационных систем с использованием логических методов классификации.