Федеральное государственное образовательное бюджетное учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»
(Финансовый университет)
Колледж информатики и программирования

Тема занятия «Алгоритмы машинного обучения: Решающие деревья, ансамблевые методы»

Машинное обучение активно используется в современных IT-компаниях для решения разнообразных задач. Знание таких алгоритмов, как решающие деревья и ансамблевые методы, крайне важно для будущих программистов и аналитиков данных, открывая широкие возможности для карьерного роста.

Дисциплина: ОП.14 «Основы машинного обучения»
Специальность: 09.02.07 Информационные системы и программирование
Преподаватель: Горланов Владимир Владимирович

Цели и задачи урока

Цель

Сформировать умение выбирать и применять алгоритмы машинного обучения для решения типовых задач.

Задачи

- Актуализировать знания о базовых понятиях машинного обучения.
- Изучить принципы работы и применения решающих деревьев, ансамблевых методов

Формируемые общепрофессиональные компетенции (ОК)

- ОК 01 Выбирать способы решения задач профессиональной деятельности применительно к различным контекстам.
- ОК 02 Осуществлять поиск, анализ и интерпретацию информации, необходимой для выполнения задач профессиональной деятельности.
- ОК 09 Использовать информационные технологии в профессиональной деятельности.

Онлайн-викторина

Изучение нового материала
Задание: заполнить таблицу в ходе лекции

Таблица «Знаю — Хочу узнать — Узнал»

Знаю	Хочу узнать	Узнал
Что уже знаете по теме	Формулируйте вопрос, что хотите узнать	Что узнали после изучения темы

План лекции

1

Решающие деревья

Структура, построение и применение

2

Ансамблевые методы

Стекинг, бэггинг, бустинг: особенности и примеры

3

Практическое задание

Отработка полученных знаний в групповой работе

Решающее дерево

Рассмотрим один из популярных алгоритмов — **решающее дерево**. Оно может описывать процесс принятия решения почти в любой ситуации.

Пример: премия сотруднику

Проверяем условия по порядку:

- Выполнял ли все задачи в срок?
- Предлагал ли инновационные идеи?
- Если условия выполнены — получит премию

Энтропия

Энтропия — важное понятие из физики и теории информации. Интуитивно энтропия соответствует **степени хаоса в системе**.

Пример с корзиной шаров: представим корзину с красными и синими шарами.

- Сначала шары разделены — низкая энтропия, то есть степень хаоса низкая
- После встряхивания шары смешиваются — высокая энтропия, то есть степень хаоса высокая

Алгоритм построения дерева

Изначально классы в данных перемешаны. Цель алгоритма — найти решающее дерево, которое отделит их друг от друга. Правила разделения строятся так, чтобы **уменьшать степень хаоса (энтропию)** в выборке.

1

Высокая степень хаоса (энтропия)

Маленький прирост информации

2

Низкая степень хаоса (энтропия)

Большой прирост информации

Алгоритм использует принцип жадной максимизации: на каждом шаге выбирается признак с наибольшим приростом информации. Процедура повторяется рекурсивно до достижения минимальной энтропии.

Обучение решающего дерева

Как выбрать лучшее из всего возможного множества деревьев? Чтобы найти оптимальное решение, нужно обучить модель: подобрать решающее дерево, которое больше всего подойдёт нашей обучающей выборке.

1 **Алгоритмы обучения**

Отвечают за поиск оптимального дерева

2 **Обученная модель**

Результат работы алгоритмов обучения

3 **Предсказания**

Модель получает новые объекты и выдаёт ответы

После обучения модель готова работать самостоятельно — больше не нужны алгоритмы и обучающий набор данных.

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Пример построения дерева. Постановка задачи

Построим дерево для определения, есть ли у ученика тройки в табеле, по трём бинарным признакам. Обучаемся на данных о шести учениках.

Изначально все ученики находятся в одном узле с максимальной степенью хаоса — поровну представителей каждого класса (3 и 3).

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Первое решающее правило

Выберем первый признак для разделения учеников, анализируя прирост информации:

«Все ли уроки посещает»

«Делает ли все ДЗ»

«Посещает ли олимпиады»

Посещает ли олимпиады?

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Кристина	0	1	1	1

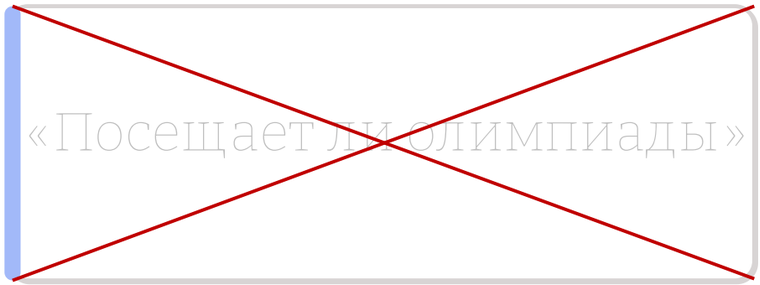
Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Первое решающее правило

Выберем первый признак для разделения учеников, анализируя прирост информации:

«Все ли уроки посещает»

«Делает ли все ДЗ»



Делает ли все ДЗ?

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Света	0	0	0	0

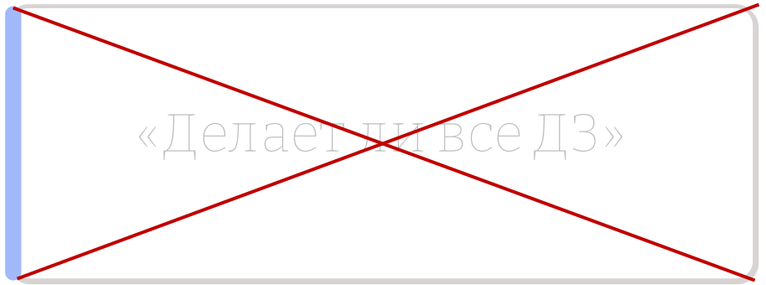
Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1

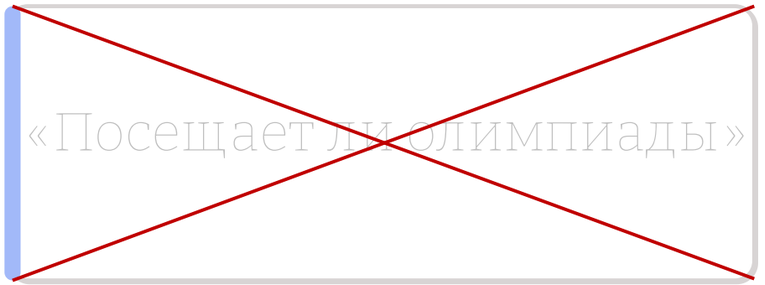
Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Первое решающее правило

Выберем первый признак для разделения учеников, анализируя прирост информации:

«Все ли уроки посещает»





Все ли уроки посещает?

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Света	0	0	0	0

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Егор	1	1	0	1

Ученик	Все ли уроки посещает	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Вася	1	0	0	1
Петя	0	0	0	0
Кристина	0	1	1	1
Полина	0	1	0	0
Егор	1	1	0	1
Света	0	0	0	0

Первое решающее правило

Выберем первый признак для разделения учеников, анализируя прирост информации:

«Все ли уроки посещает»

В одном листе — только ученики без троек, в другом — небольшой хаос (3 троечника и 1 без троек)

«Делает ли все ДЗ»

В каждом листе остаются перемешанными троечники с учениками без троек

«Посещает ли олимпиады»

Слабо уменьшает хаос — среди непосещающих похожее число троечников и учеников без троек

✓ Выбираем разделение по признаку «Все ли уроки посещает» — максимальный прирост информации

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0
Кристина	1	1	1
Полина	1	0	0
Света	0	0	0

Второе решающее правило

В одном листе (посещающие все уроки) хаос нулевой — делить не нужно.
Во втором листе: 3 троечника и 1 ученик без троек.

- «Делает ли все ДЗ»
- «Посещает ли олимпиады»

Делает ли все ДЗ?

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0
Света	0	0	0

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Кристина	1	1	1
Полина	1	0	0

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0
Кристина	1	1	1
Полина	1	0	0
Света	0	0	0

Второе решающее правило

В одном листе (посещающие все уроки) хаос нулевой — делить не нужно.
Во втором листе: 3 троечника и 1 ученик без троек.

- ~~«Делает ли все ДЗ»~~
- «Посещает ли олимпиады»

Посещает ли олимпиады?

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0
Полина	1	0	0
Света	0	0	0

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Кристина	1	1	1

Ученик	Делает ли все ДЗ	Посещает ли олимпиады	Не имеет троек в табеле
Петя	0	0	0
Кристина	1	1	1
Полина	1	0	0
Света	0	0	0

Второе решающее правило

В одном листе (посещающие все уроки) хаос нулевой — делить не нужно.
Во втором листе: 3 троечника и 1 ученик без троек.

«Делает ли все ДЗ»

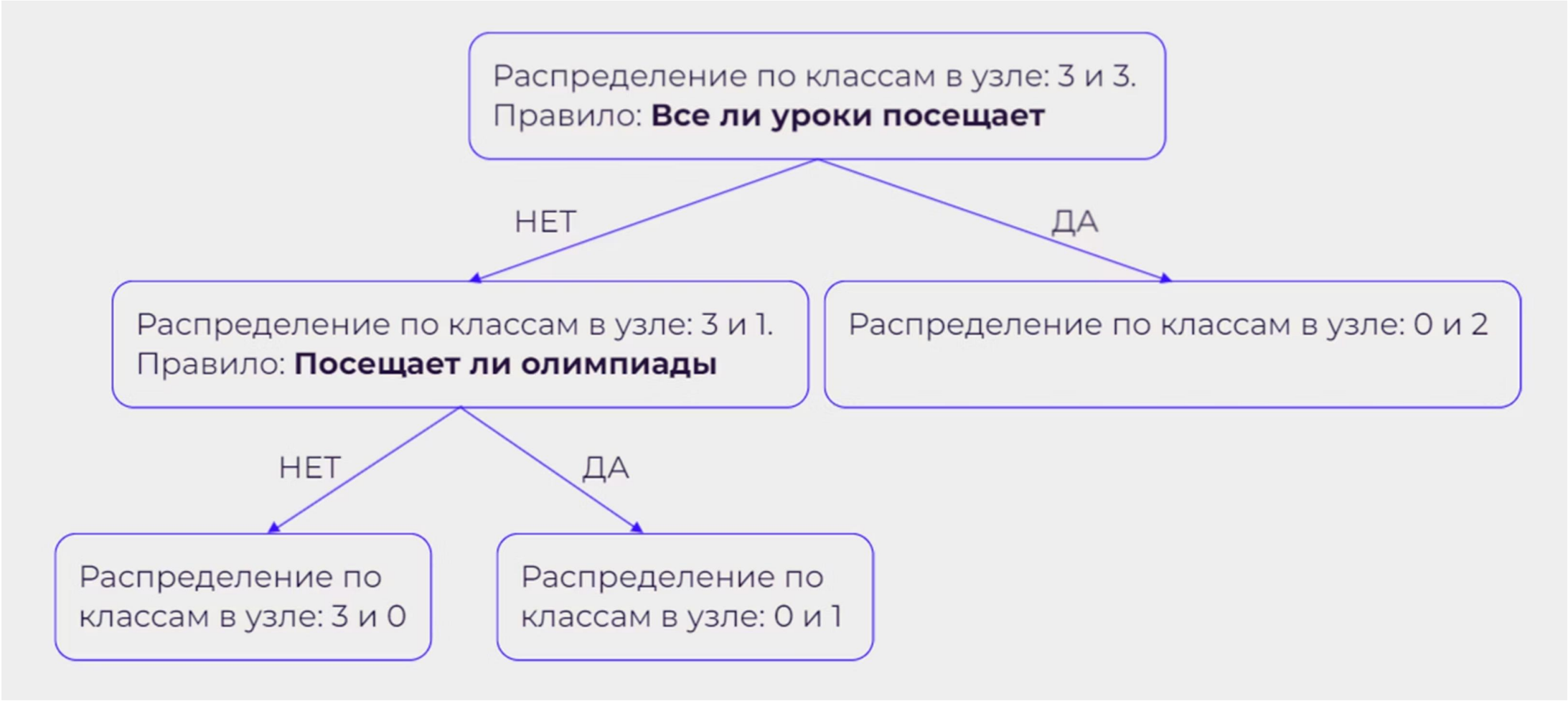
Один лист — только троечники, другой — 1 троечник и 1 без троек. Хаос остаётся

«Посещает ли олимпиады»

Один лист — все троечники, другой — ученица без троек. Минимальная степень хаоса

Выбираем разделение по признаку «Посещает ли олимпиады»

Итоговое дерево



Применение решающих деревьев

Медицинская диагностика

Правила принятия решений на основе симптомов пациента для диагностики заболеваний

Кредитный рейтинг

Принятие решения о выдаче кредита по набору финансовых и личных характеристик заемщика

Бизнес-аналитика

Анализ клиентского поведения, сегментация рынка и прогнозирование продаж

В **Scikit-Learn** используйте `DecisionTreeClassifier` для классификации и `DecisionTreeRegressor` для регрессии.

Ансамбли алгоритмов

Ансамблевый метод — это подход машинного обучения, где несколько моделей обучаются для решения одной проблемы и объединяются для получения лучших результатов.

Принцип работы

Ансамбль работает лучше, если базовые модели **разнородны по ошибкам** – каждая ошибается по-разному, объединение сглаживает индивидуальные промахи.

Множество простых моделей может превзойти одну сложную модель за счёт усреднения и компенсации ошибок.

Слабый ученик

Одна простая модель (например, дерево решений)

Сильный ученик

Объединение слабых учеников для улучшения качества

Виды ансамблевых методов

Наиболее популярными ансамблевыми методами являются: **стекинг, бэггинг, бустинг**.

Стекинг

Использует несколько разнородных слабых учеников. Их обучают и объединяют для построения прогноза на основе различных моделей.

Аналогия: перед принятием итогового решения, вы советуетесь с разными людьми из разных областей, после чего уже принимаете итоговое решение.

Бэггинг

Однородные модели обучают на разных наборах данных и объединяют. Прогноз получают путём усреднения.

Аналогия: 10 экспертов решают задачу на основе своего опыта, затем голосуют. Общий вердикт надёжнее мнения одного.

Бустинг

Несколько однородных моделей последовательно обучаются, исправляя ошибки друг друга.

Аналогия: готовитесь к экзамену, находите ошибки, доучиваете слабые места, снова тестируете – итоговый уровень намного выше.

Стекинг

Работа стекинга довольно проста:

01

Слабые предсказатели

На вход подаётся обучающий набор

02

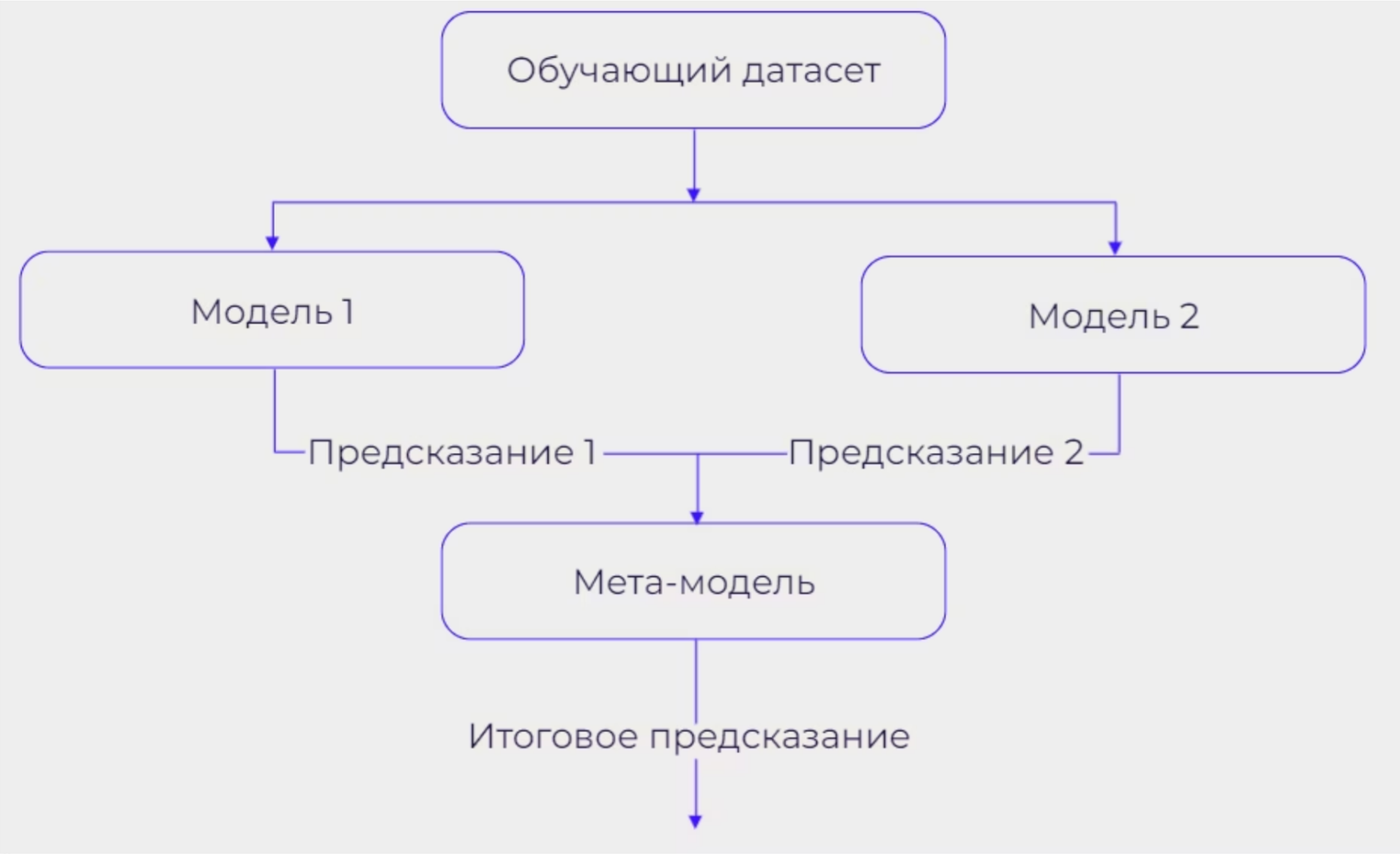
Мета-модель

Каждый прогноз идёт к финальной модели (смеситель)

03

Финальный прогноз

Мета-модель вырабатывает итоговое решение



Бэггинг

Основная идея **бэггинга** — обучить несколько одинаковых моделей на разных образцах. Поскольку распределение выборки неизвестно, модели получатся разными.

Бутстрэп-выборки

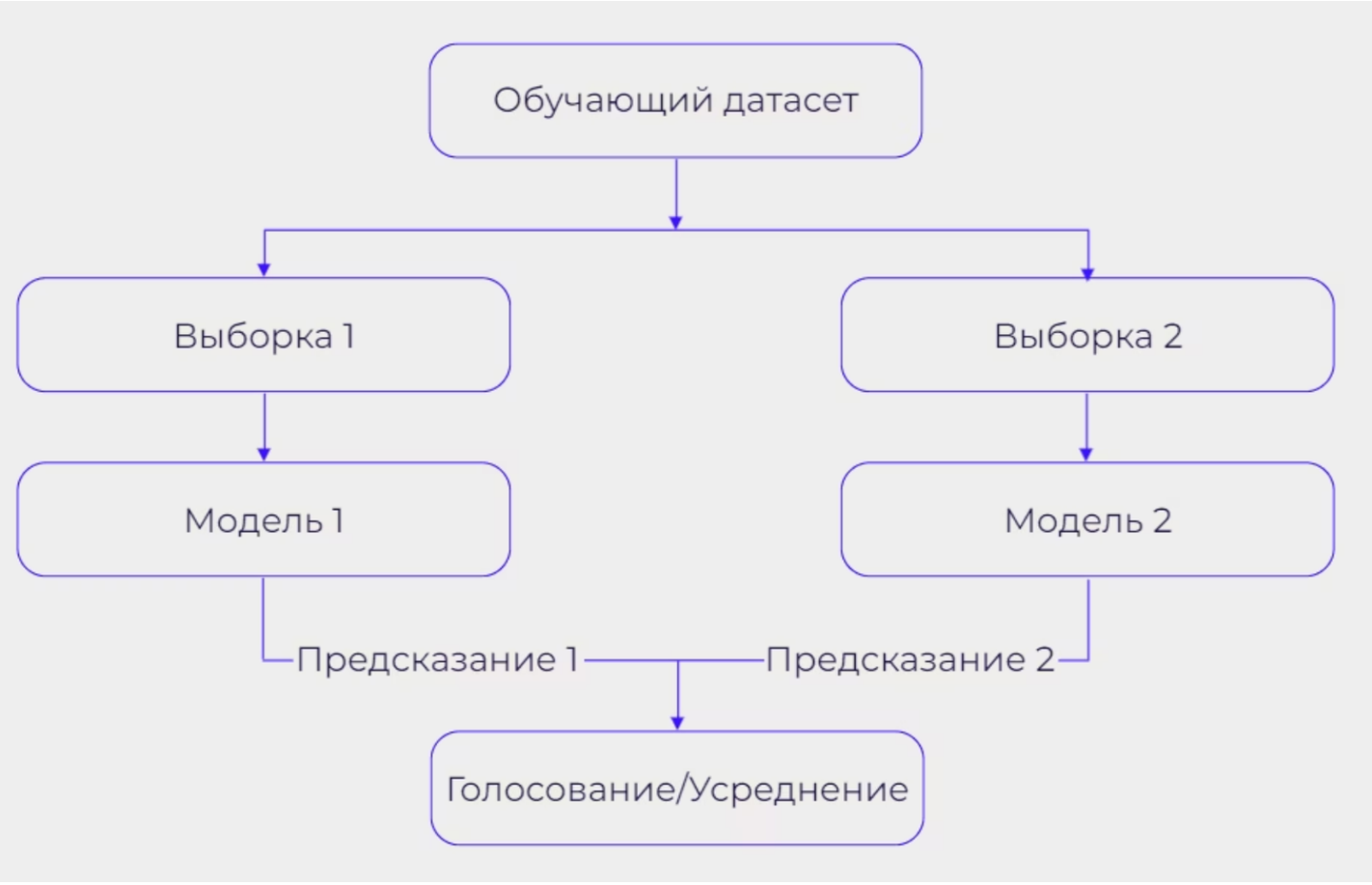
Случайный выбор данных из датасета с возвращением

Обучение моделей

Каждая модель обучается на своей выборке

Объединение прогнозов

Усреднение для регрессии, голосование для классификации



Типы голосования в бэггинге

Жёсткое голосование

Класс, который предсказывает большинство слабых моделей, получает больше голосов и становится результатом ансамбля

Мягкое голосование

Рассматриваются вероятности предсказания каждого класса, затем вероятности усредняются и результатом является класс с большой вероятностью

Преимущество бэггинга: обучение можно распараллелить, поскольку модели не зависят друг от друга

Случайный лес

Самый популярный пример **бэггинга** — алгоритм случайного леса. Он обучает большое количество независимых друг от друга деревьев, а потом принимает решение на основе голосования.

Случайные данные

Для каждого дерева берется случайный набор данных (бутстрэп-выборки)

Случайные критерии

Каждое дерево использует случайный набор признаков для принятия решений

Случайным он называется именно из-за этой двойной случайности в выборе данных и критериев.

Бустинг

Метод **бустинга** схож с бэггингом: берётся множество одинаковых моделей и объединяется для получения сильного учени

Последовательное обучение

Модели приспосабливаются к данным последовательно, каждая исправляет ошибки предыдущей

Зависимость моделей

Каждая новая модель зависит от предыдущей, поэтому распараллелить обучение нельзя

Это ключевое отличие от бэггинга, где модели обучаются независимо и параллельно.

Практическое задание

Выполните практическое задание по группам, используя полученные знания по алгоритмам машинного обучения.

Основная задача:
Интернет-магазин хочет внедрить систему персонализированных рекомендаций для увеличения продаж и улучшения пользовательского опыта.

1

Распределение алгоритмов

- **Группа 1-2:** Решающее дерево
- **Группа 3-4:** Стекинг
- **Группа 5-6:** Случайный лес
- **Группа 7-8:** Бустинг
- **Группа 9-10:** Бэггинг

2

Роли в группе (3 человека)

- **Аналитик данных:** анализирует датасет и выявляет закономерности.
- **ML-инженер:** выбирает и настраивает алгоритмы.
- **Бизнес-консультант:** интерпретирует результаты для бизнеса.

Требования к мини-презентации (2 минуты):

- Распределение ролей в вашей группе.
- Краткое представление алгоритма, который был дан вашей группе.
- Демонстрация результатов обучения вашей модели.

Критерии оценивания практической работы

Для успешного выполнения практического задания и получения высокой оценки, пожалуйста, уделите внимание следующим аспектам:

Сотрудничество в группе

Эффективное распределение и выполнение обязанностей, активное участие в командной работе.

Поведение

Соблюдение правил работы в аудитории, не отвлекаться от выполнения задания и не мешать другим группам.

Раскрытие материала и задания

Глубокое понимание и применение теоретических знаний, полное выполнение всех частей задания.

Умение слушать и задавать вопросы

Активное участие в обсуждениях, внимательное отношение к выступлениям других групп и конструктивные вопросы.

Ваша работа будет оценена по следующей шкале:

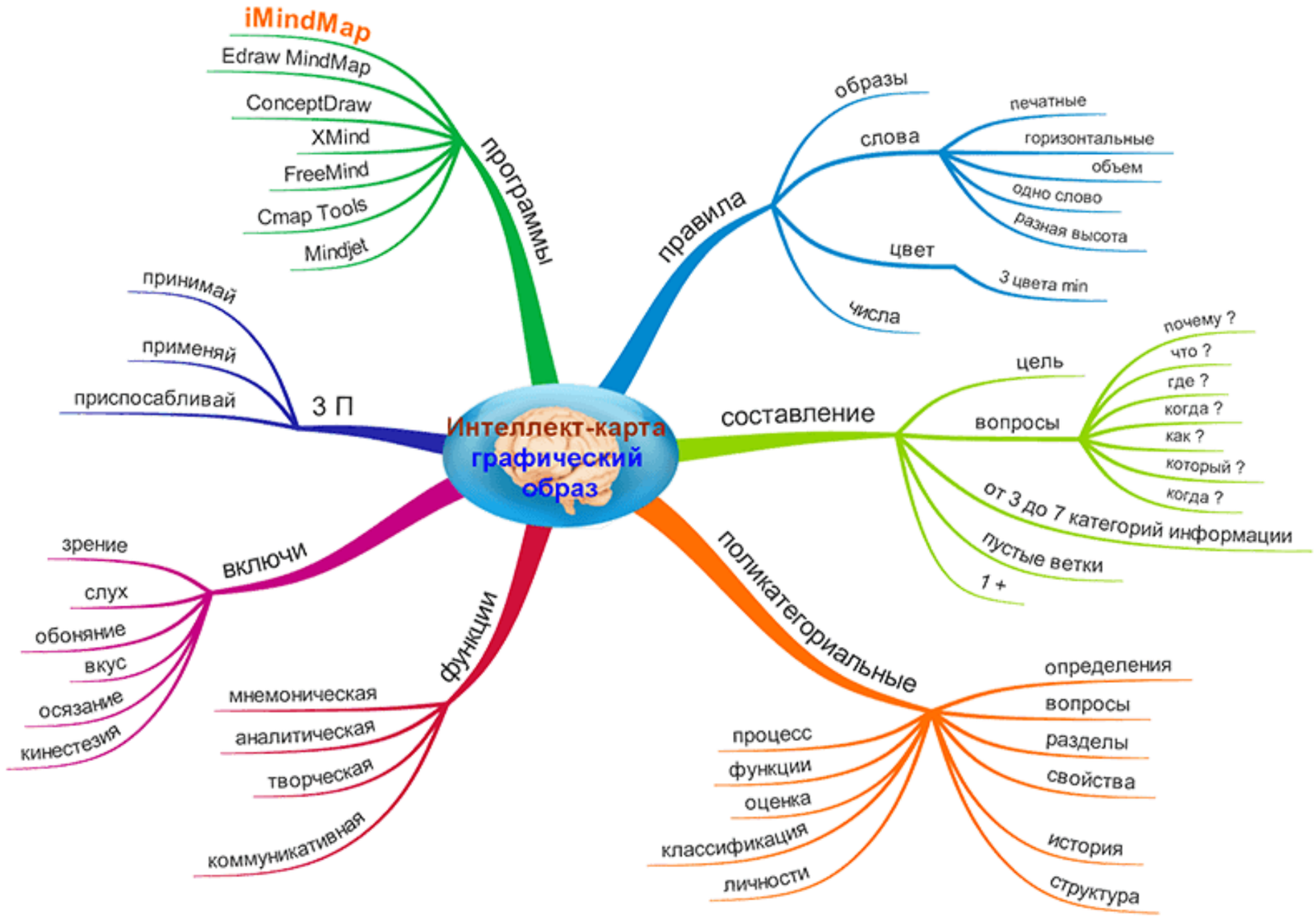
Оценка	Описание
Отлично (100-80%)	Студент выполнил все задания на высшем уровне. Проведён корректный анализ данных и написан корректный код для обучения. Результаты представлены ясно и в корректном виде. Проявлена инициатива и помощь другим студентам.
Хорошо (79-70%)	Студент выполнил большую часть заданий на высоком уровне, но допустил незначительные ошибки. Анализ данных проведён с небольшими недочетами, в коде есть незначительные ошибки. Результаты представлены в корректном виде.
Удовлетворительно (69-50%)	Студент выполнил основные задания, но допустил несколько ошибок. Проведён некорректный анализ данных, код выполнен с огрехами. Результаты представлены в корректном виде.
Неудовлетворительно (менее 50%)	Студент не выполнил значительную часть заданий или допустил критические ошибки. Проведён некорректный анализ данных. Написан некорректный код, либо код не был написан. Результаты не представлены.

Подведение итогов, рефлексия

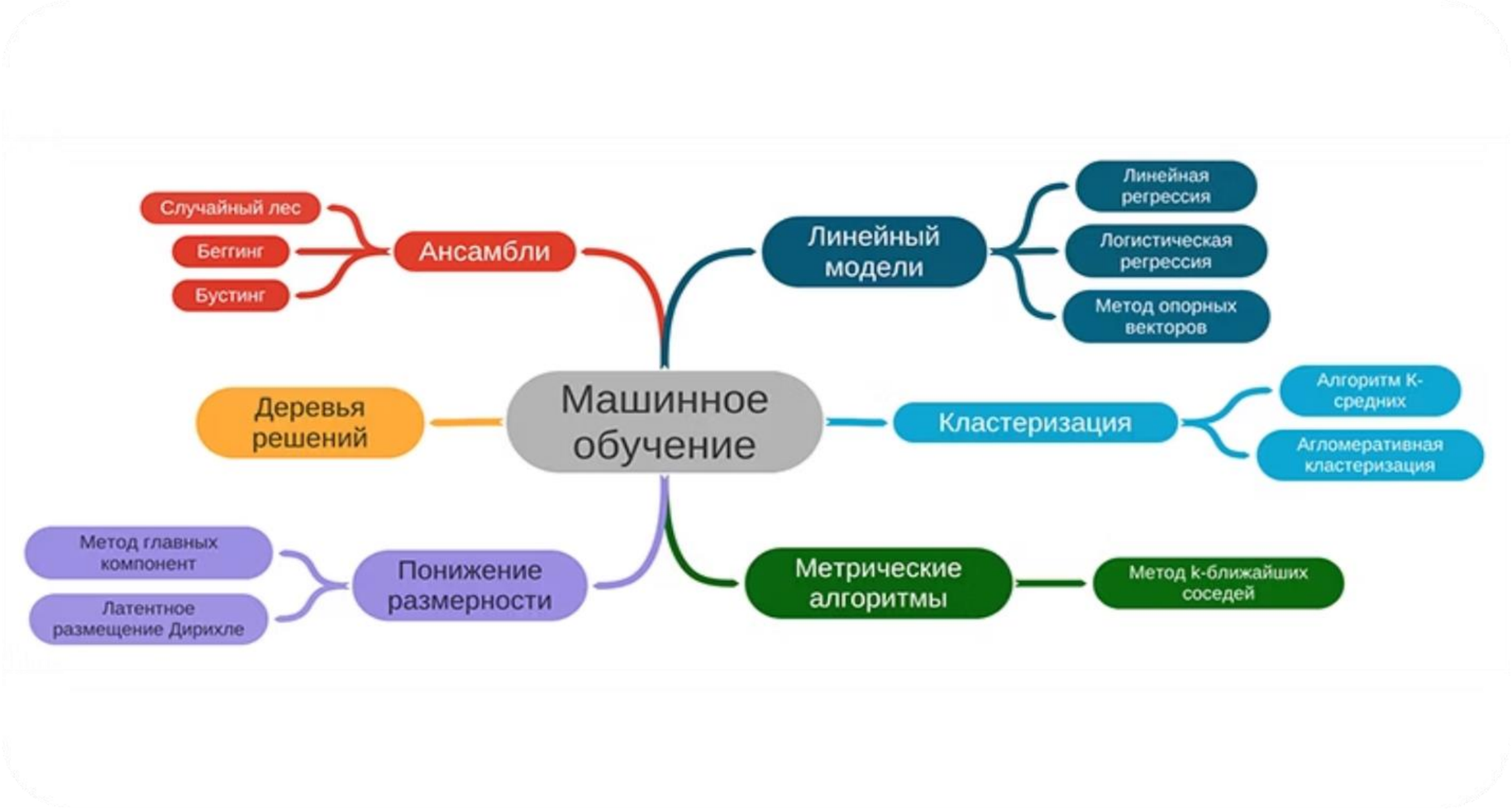
Пожалуйста, уделите несколько минут для самоанализа и ответьте на следующие вопросы:

1. Хватило ли вам теоретических знаний на занятии для того, чтобы понять, как реализовать задачу?
2. Были ли какие-то моменты на лекции, которые показались особенно сложными или непонятными?
3. Как вы себя чувствуете, уверены ли вы в своих силах?
4. Получилась ли у вас работа на практике, или возникли какие-то трудности в реализации кода? Какие именно?
5. Какие части занятия были для вас наиболее интересными, а какие – сложными?
6. Если бы вы могли что-то улучшить в своем обучении на этом занятии, чтобы это было?

Задание: оформите интеллект-карту по изученной теме



Интеллект-карта



Список литературы и электронных ресурсов:

1. Машинное обучение. – [Электронный ресурс]. – Режим доступа: https://koroteev.site/ml/

2. Учебник по машинному обучению. Решающие деревья. – [Электронный ресурс].
– Режим доступа: https://education.yandex.ru/handbook/ml/article/reshayushchiye-derevya

3. Учебник по машинному обучению. Ансамбли в машинном обучении. – [Электронный ресурс].
– Режим доступа: https://education.yandex.ru/handbook/ml/article/ansambli-v-mashinnom-obuchenii

4. Основы машинного обучения: учебник / М.В. Кортеев. – Москва: Кинорус, 2024. – 431 с.
– (Высшее образование: Бакалавриат). – ISBN 978-5-406-12673-8. – Текст: электронный.
– Режим доступа: https://book.ru/book/952751