ПРИЛОЖЕНИЕ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ
В.Ф. УТКИНА»

Кафедра «Электронные вычислительные машины»

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ

Б1.В.ДВ.01.01 «Интеллектуальный анализ больших данных»

Направление подготовки
02.03.03 Математическое обеспечение и администрирование
информационных систем

Профиль
«Программное обеспечение компьютерных технологий и систем
искусственного интеллекта»

Уровень подготовки
Бакалавриат

Квалификация выпускника – бакалавр

Форма обучения – очная

Рязань 2025

1. ОБЩИЕ ПОЛОЖЕНИЯ

Оценочные материалы – это совокупность учебно-методических материалов (контрольных заданий, описаний форм и процедур проверки), предназначенных для оценки качества освоения обучающимися данной дисциплины как части ОПОП.

Цель – оценить соответствие знаний, умений и владений, приобретенных обучающимся в процессе изучения дисциплины, целям и требованиям ОПОП в ходе проведения промежуточной аттестации.

Промежуточный контроль по дисциплине осуществляется путем проведения экзамена. Форма проведения экзамена – билеты с письменным ответом на два теоретических вопроса и одним практическим заданием. При необходимости, проводится устная беседа с обучаемым для уточнения оценки. Выполнение заданий на практических занятиях в течение семестра и заданий на самостоятельную работу является обязательным условием для допуска к экзамену.

2. ПАСПОРТ ОЦЕНОЧНЫХ МАТЕРИАЛОВ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Контролируемые разделы (темы) дисциплины (результаты по разделам)	**Код контролируемой компетенции (или её части)**	**Наименование оценочного средства**
Раздел 1. Основные понятия и концепции анализа больших данных.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 2. Архитектура хранилищ и платформ хранения данных.	ПК-8, ПК-13, ПК-21	Экзамен
Раздел 3. Статистический анализ и визуализация данных.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 4. Машинное обучение и методы классификации и кластеризации.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 5. Глубокое обучение и нейронные сети.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 6. Распределённые вычисления и параллельные алгоритмы.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 7. Интеграция и преобразование данных.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 8. Анализ поведения пользователей и персонализация услуг.	ПК-8, ПК-10, ПК-13, ПК-21	Экзамен
Раздел 9. Прогнозирование и принятие решений.	ПК-8, ПК-10, ПК-21	Экзамен

3. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ ОЦЕНИВАНИЯ КОМПЕТЕНЦИЙ

Сформированность компетенции (или ее части) в рамках освоения данной дисциплины

оценивается по трехуровневой шкале:

1) Пороговый (базовый) уровень является обязательным для всех обучающихся по завершении освоения дисциплины;

2) Продвинутый уровень характеризуется превышением минимальных характеристик сформированности компетенций по завершении освоения дисциплины;

3) Эталонный (экспертный*)* уровень характеризуется освоением компетенций на уровне выше среднего и является важным качественным ориентиром для самосовершенствования.

Уровень освоения компетенций, формируемых дисциплиной:

Описание критериев и шкалы оценивания экзаменационного билета:

Шкала оценивания	Критерий
5 баллов (эталонный уровень)	выставляется студенту, который дал полные ответы на вопросы, показал глубокие систематизированные знания, смог привести примеры, решил практическую задачу, ответил на дополнительные вопросы преподавателя
4 балла (продвинутый уровень)	выставляется студенту, который дал преимущественно полные ответы на вопросы, решил практическую задачу, но на некоторые дополнительные вопросы преподавателя ответил только с помощью наводящих вопросов
3 балла (пороговый уровень)	выставляется студенту, который дал неполные ответы на вопросы в билете, показал в основном верный ход решения задачи и смог ответить на дополнительные вопросы только с помощью преподавателя
2 балла	выставляется студенту, который не смог ответить на вопросы, а также решить практическую задачу

4. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ

4.1. Промежуточная аттестация

Коды компетенций	Результаты освоения ОПОП Содержание компетенций (код и содержание индикатора)
ПК-8	***Способен применять язык программирования Python для решения задач в области ИИ*** ПК-8.1. Разрабатывает и отлаживает прикладные решения разной сложности и для разного круга конечных пользователей с использованием языка программирования Python, тестирует, испытывает и оценивает качество таких решений. ПК-8.2. Осуществляет выбор инструментов разработки на Python, приемлемых для создания прикладной системы обработки научных данных, машинного обучения и визуализации с заданными требованиями. ПК-8.3. Разрабатывает и поддерживает системы обработки больших данных различной степени сложности.
ПК-10	***Способен осуществлять поиск сбор очистку и предварительный анализ***

	данных ПК-10.1. Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи. ПК-10.2. Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ.
ПК-13	***Способен применять различные модели и (или) технологии обработки данных*** ПК-13.1. Осуществляет выбор технологий обработки больших данных, приемлемых для создания прикладной системы ИИ с заданными требованиями. ПК-13.2. Разрабатывает и отлаживает прикладные решения с элементами ИИ с применением различных технологий обработки данных.
ПК-21	***Способен применять современную теоретическую математику для разработки новых алгоритмов и формулирования перспективных задач ИИ*** ПК-21.1. Обосновывает способы и варианты применения методов и моделей в задачах искусственного интеллекта, включая их модификацию и адаптацию кспецифике задачи. ПК-21.2. Применяет аппарат теории вероятностей, матстатистики и теории информации для формулирования и анализа задач искусственного интеллекта.

Тестовые вопросы для оценки уровня освоения компетенций:

Тема 1 «Основные понятия и концепции анализа больших данных»

Вопрос №1

Какой минимальный объем данных классифицируется как «большие данные»?
Варианты:
a) Несколько десятков мегабайт
b) Сотни гигабайт
c) Десять гигабайт
d) Несколько сотен килобайт
e) Один террабайт

Вопрос №2

В чём основное отличие больших данных от традиционных данных?
Варианты:
a) Малый объем данных
b) Отсутствие специального оборудования для обработки
c) Потребность в особой инфраструктуре для хранения и обработки
d) Единая схема хранения данных
e) Однотипность формата данных

Вопрос №3

Кто первым ввёл термин «большие данные»?
Варианты:
a) Даглас Энгельбарт

b) Марк Цукерберг
c) Клиффорд Линч
d) Стив Джобс
e) Билл Гейтс

Вопрос №4
Какие типы данных включаются в характеристику «разнообразие» (variety)?
Варианты:
a) Только структурированные данные
b) Структурированные, неструктурированные и частично структурированные данные
c) Только неструктурированные данные
d) Только текстовая информация
e) Только графические данные

Вопрос №5
Что означает характеристика «скорость» (velocity) в контексте Big Data?
Варианты:
a) Скорость изменения погоды
b) Число серверов для обработки данных
c) Частота пополнения и обновления данных
d) Длительность жизни товара
e) Объём хранимой информации

Вопрос №6
Верно ли утверждение, что Big Data начали использоваться ранее 2011 года?
Варианты:
a) Да, научные исследования проводились ещё в XX веке
b) Нет, начало использования связано с 2012 годом
c) Первое широкое применение было зафиксировано в 2014 году
d) Появилось только в первой половине 2000-х годов
e) Впервые возникло в 2020-е годы

Вопрос №7
Какие отрасли экономики активно применяют Big Data сегодня?
Варианты:
a) Образование и наука
b) Банковская сфера, медицина, ритейл, промышленное производство
c) Спортивные мероприятия и туризм
d) Строительство и дизайн интерьеров
e) Искусство и кинематограф

Вопрос №8
Из каких компонентов состоит экосистема Hadoop?
Варианты:
a) PHP и CSS
b) MapReduce, YARN, HDFS
c) MySQL и SQLite
d) HTML и JavaScript
e) ASP.NET и ColdFusion

Вопрос №9

Какие технологии поддерживают обработку данных в реальном времени?

Варианты:

a) WordPress и Joomla

b) Photoshop и Illustrator

c) Apache Spark и Apache Storm

d) AutoCAD и Revit

e) MS Office и LibreOffice

Вопрос №10

Что подразумевается под демократизацией данных?

Варианты:

a) Ограничение доступа сотрудников к аналитическим ресурсам

b) Предоставление сотрудникам доступа к аналитическим инструментам вне IT-подразделений

c) Снижение квалификации сотрудников для упрощения доступа к данным

d) Укрепление административного контроля над аналитическими средствами

e) Ужесточение мер безопасности при доступе к данным

Вопрос №11

Что означает характеристика «точность» (veracity) в характеристиках Big Data?

Варианты:

a) Цена обработки данных

b) Качество подготовки отчетов

c) Степень доверия к данным и результатам их анализа

d) Величина прироста прибыли от данных

e) Средняя продолжительность хранения данных

Вопрос №12

Что характеризует вариативность (variability) в контексте Big Data?

Варианты:

a) Устойчивость данных

b) Изменчивость и колебания в потоке данных

c) Привязанность данных к географическому расположению центра обработки

d) Стабильность функционирования сети

e) Частота появления новых данных

Вопрос №13

Что понимается под характеристикой «ценность» (value) в описании Big Data?

Варианты:

a) Стоимостные показатели закупки серверов

b) Показатель популярности поисковых запросов

c) Важность данных для принятия решений и проведения анализа

d) Срок годности технических устройств

e) Среднее время ожидания обработки данных

Вопрос №14

Какие трудности возникают при внедрении Big Data в компании?

Варианты:

a) Высокие издержки на технику и подготовку специалистов

b) Быстродействие транспортных перевозок
c) Сложности выбора подходящей технологии
d) Необходимость модернизации устаревшей инфраструктуры
e) Все перечисленные варианты

Вопрос №15

Какие законодательные нормы регулируют обращение с персональными данными?

Варианты:

a) Международные соглашения
b) Законы о защите прав потребителей
c) В международном праве это GDPR, HIPAA, CCPA, в РФ это 152 ФЗ «О защите персональных данных»
d) Законы о защите авторских прав
e) Налоговый кодекс

Вопрос №16

Какие тенденции ожидаются в развитии Big Data в ближайшее время?

Варианты:

a) Переход обратно к бумажным архивам
b) Внедрение технологий обработки данных в реальном времени и облачных сервисов
c) Забвение аналитики данных
d) Возврат к доминированию физических носителей информации
e) Ослабление значимости аналитических инструментов

Вопрос №17

Какие будущие направления выделяет большинство экспертов относительно Big Data?

Варианты:

a) Падение значимости аналитики
b) Расширение использования продвинутых аналитических методик
c) Отказ от цифровизации процессов
d) Сокращение финансирования инновационных разработок
e) Увеличение объёмов ручной обработки данных

Вопрос №18

Какова цель демократизации данных?

Варианты:

a) Установление запретов на доступ сотрудников к данным
b) Увеличение продуктивности сотрудников путём предоставления доступа к аналитическим средствам
c) Добавление бюрократических препятствий
d) Удорожание стоимости данных
e) Исчезновение необходимости в обучении сотрудников

Вопрос №19

Какая главная проблема возникает при хранении и обработке больших данных?

Варианты:

a) Недоступность флеш-накопителей большой ёмкости
b) Непригодность традиционного ПО для эффективного использования данных

c) Высокая вероятность случайной потери данных
d) Трудности масштабирования и распределения ресурсов
e) Наличие множества бесплатных инструментов для обработки данных

Вопрос №20
Какую роль играет аналитика больших данных в устойчивом развитии?
Варианты:
a) Поддерживает повышение энергоэффективности и снижение выбросов углекислого газа
b) Приводит к росту энергопотребления предприятий
c) Усугубляет проблему утилизации электронных отходов
d) Снижает уровень переработки материалов
e) Препятствует разработке новых экологически чистых технологий

Вопрос №21
Какие технологии используются для анализа больших данных в реальном времени?
Варианты:
a) MySQL и PostgreSQL
b) Apache Kafka и Apache Spark
c) Notepad++ и Visual Studio Code
d) Sublime Text и Atom Editor
e) SketchUp и Blender

Вопрос №22
Какие преимущества предоставляют облачные сервисы для обработки больших данных?
Варианты:
a) Возможность играть в компьютерные игры с высоким разрешением
b) Экономическая выгода благодаря оплате по мере использования ресурса
c) Упрощённое оформление интерьера офисов
d) Незаменимая помощь в трудоустройстве
e) Бесплатные образовательные программы

Вопрос №23
Какие типы баз данных предпочтительнее для работы с большими данными?
Варианты:
a) Реляционные базы данных SQL
b) Базы данных NoSQL
c) Локальные папки с файлами CSV
d) Таблицы Excel
e) Блокнотные записи на бумаге

Вопрос №24
Какие шаги важны для успеха проекта по обработке больших данных?
Варианты:
a) Активная покупка дорогого оборудования и привлечение дизайнеров
b) Повышение квалификации сотрудников
c) Ясное формулирование целей и выделение достаточных ресурсов
d) Сокращение штата сотрудников
e) Ограничение доступа сотрудников к аналитическим материалам

Вопрос №25

Какое преимущество даёт организациям использование анализа больших данных?

Варианты:

a) Быстрое выявление нарушений и рисков

b) Невозможность получения инсайтов из накопленной информации

c) Замедление процессов выпуска новых продуктов

d) Утрата контакта с клиентурой

e) Дополнительные расходы на администрирование данных

Вопрос №26

Какие проблемы связаны с безопасностью и конфиденциальностью больших данных?

Варианты:

a) Малая активность злоумышленников

b) Недостаточный уровень профессионализма специалистов и сложность противодействия угрозам

c) Открытый доступ пользователей ко всем видам данных

d) Самостоятельное исправление всех возникших проблем системами

e) Широкий выбор готовых инструментов для устранения угроз

Вопрос №27

Какова основная задача аналитики больших данных?

Варианты:

a) Создание красивых презентаций

b) Формулировка гипотез без подтверждения фактов

c) Выявление скрытых зависимостей и формирование рекомендаций

d) Генерация дополнительных отчётов

e) Произвольное удаление элементов данных

Вопрос №28

Какие плюсы несет переход данных в облачные среды?

Варианты:

a) Удорожание эксплуатационных расходов

b) Возможности гибкого масштабирования и экономии средств

c) Обязательная замена имеющегося оборудования

d) Потеря контроля над собственными данными

e) Затруднения в совместном использовании данных разными подразделениями

Вопрос №29

Какие главные вызовы стоят перед индустрией Big Data прямо сейчас и в ближайшей перспективе?

Варианты:

a) Утрата ценности данных

b) Защита данных и усиление конфиденциальности

c) Тенденция отказа от цифровых технологий

d) Растущие дефициты компьютерной грамотности среди сотрудников

e) Совершенствование методов физического хранения данных

Тема 2 «Архитектура хранилищ и платформ хранения данных»

Вопрос №1

Что такое распределённая система?

Варианты:

a) Система, выполняемая на одном компьютере

b) Набор независимых компьютеров, воспринимаемый пользователями как единое целое

c) Процессор с несколькими ядрами

d) Сеть общего пользования

e) Сервер с поддержкой удалённого доступа

Вопрос №2

Назовите основной принцип горизонтальной масштабируемости (scale-out).

Варианты:

a) Увеличение количества дисков на существующем сервере

b) Добавление новых узлов в систему

c) Увеличение тактовой частоты процессора

d) Увеличение объёма оперативной памяти

e) Апгрейд сетевого оборудования

Вопрос №3

Что характерно для отказоустойчивости распределённой системы?

Варианты:

a) Синхронизация часов на всех узлах

b) Централизованное хранилище журналов

c) Репликация данных и резервирование компонентов

d) Совместная работа с одним общим диском

e) Централизованная аутентификация пользователей

Вопрос №4

Что означает CAP-теорема?

Варианты:

a) Теория о выборе оптимального процессора

b) Компьютерная программа для мониторинга трафика

c) Три аспекта распределённых систем: согласованность, доступность и устойчивость к разделению

d) Метод шифрования данных

e) Алгоритм сжатия данных

Вопрос №5

Как называется свойство распределённой системы, гарантирующее согласованность данных?

Варианты:

a) Availability (Доступность)

b) Consistency (Согласованность)

c) Partition Tolerance (Устойчивость к разделению)

d) Elasticity (Эластичность)

e) Latency (Задержка)

Вопрос №6

Какой режим масштабирования поддерживает быстрое добавление новых узлов?

Варианты:
a) Vertical Scale (Вертикальная масштабируемость)
b) Horizontal Scale (Горизонтальная масштабируемость)
c) Manual Scale (Ручная настройка)
d) Dynamic Scale (Динамическая балансировка нагрузки)
e) Hybrid Scale (Гибридная настройка)

Вопрос №7

Какая система гарантирует высокую доступность за счёт жертвования согласованностью?

Варианты:
a) ACID-системы
b) BASE-системы
c) RAID-массивы
d) Логические диски
e) Механизмы кеширования

Вопрос №8

Что определяет CAP-теорема как обязательное свойство распределённых систем?
Варианты:
a) Согласованность
b) Доступность
c) Устойчивость к разделению
d) Производительность
e) Безопасность

Вопрос №9

Что характеризует архитектуру HDFS (Hadoop Distributed File System)?
Варианты:
a) Централизованный сервер хранения данных
b) Блочное распределение данных с возможностью репликации
c) Одноранговый протокол передачи файлов
d) Инкапсуляция данных в бинарные пакеты
e) Клиент-серверная модель без репликации

Вопрос №10

Что обеспечивает мастер-нода (NameNode) в HDFS?
Варианты:
a) Прямой доступ к физическим данным
b) Управление метаданными файлов и блоков
c) Выполнение вычислительных задач
d) Шифрование данных
e) Хэширование файлов

Вопрос №11

Как осуществляется размещение реплик в HDFS?
Варианты:
a) Случайным образом
b) По специальному алгоритму с учётом расположения узлов
c) Пользователь задаёт расположение вручную
d) Все реплики размещаются на одном узле

e) Каждая реплика сохраняется отдельно на диске

Вопрос №12
Что такое сплит (split) в процессе MapReduce?
Варианты:
a) Файл конфигурации
b) Временный промежуточный файл
c) Отдельный участок файла, предназначенный для обработки
d) Специальный индекс данных
e) Копия основного файла

Вопрос №13
Какая фаза выполняется сразу после Mapping в MapReduce?
Варианты:
a) Sorting (Сортировка)
b) Splitting (Разбиение)
c) Shuffling (Перестановка)
d) Reducing (Редукция)
e) Compression (Сжатие)

Вопрос №14
Что такое RDD в Apache Spark?
Варианты:
a) Редактор таблиц данных
b) Библиотека для подключения к базам данных
c) Неизменяемый распределённый набор данных
d) Распределённая операционная система
e) Скриптовый движок

Вопрос №15
Что используется для оптимизации данных в MapReduce?
Варианты:
a) Соединение двух копий мапперов
b) Применение спекулятивного исполнения
c) Повторное исполнение предыдущей фазы
d) Загрузка данных из внешней базы
e) Повторная запись всех данных

Вопрос №16
Что такое DataFrame в Apache Spark?
Варианты:
a) Физический файл данных
b) Абстрактный слой поверх RDD с поддержкой SQL
c) Библиотека для работы с графикой
d) Сервис удалённого доступа
e) Интерфейс командной строки

Вопрос №17
Как организована структура данных в Cassandra?
Варианты:
a) Relational Tables (Связанные таблицы)

b) Wide Column Store (Широкие колонковые хранилища)
c) Document-Oriented DB (Документно-ориентированная база данных)
d) Graph DB (Графовая база данных)
e) Time Series DB (Временные ряды)

Вопрос №18
Как распределяются данные в MongoDB?
Варианты:
a) Поиск по диапазону значений ключа
b) Хеширование ключа и деление на фрагменты
c) Последовательное размещение записей
d) Равномерное распределение по узлам
e) Дерево двоичного поиска

Вопрос №19
Что показывает формула QPS в колоночной СУБД ClickHouse?
Варианты:
a) Задержку ввода-вывода диска
b) Размер файла журнала
c) Количество запросов в секунду
d) Время отклика сети
e) Скорость считывания данных

Вопрос №20
Что такое сущность (entity) в концепции HDFS?
Варианты:
a) Имя пользователя
b) Физический сервер
c) Операционная система
d) Запись журнала
e) Информация о файле или директории

Вопрос №21
Что такое облачные вычисления?
Варианты:
a) Вычисления в виртуальной среде
b) Технология распределённых вычислений через интернет
c) Перечень программного обеспечения серверной
d) Базовая система резервного копирования
e) Средство электронной почты

Вопрос №22
Что представляет собой модель IaaS?
Варианты:
a) Лицензионное соглашение
b) Настройка операционной системы
c) Аренда виртуальных серверов и инфраструктуры
d) Пакетное ПО для пользователей
e) Сервис для публикации статей

Вопрос №23

Какой инструмент служит для хранения объектов в AWS?
Варианты:
a) S3
b) EC2
c) Lambda
d) Route 53
e) Aurora

Вопрос №24
Что такое утилита DataProc в Yandex.Cloud?
Варианты:
a) Сервис почтовых рассылок
b) Средство визуализации данных
c) Платформа для развертывания кластеров Hadoop и Spark
d) Консольный менеджер баз данных
e) Веб-приложения для автоматизации маркетинга

Вопрос №25
Что такое кластерный режим (Fully Distributed Mode) в Hadoop?
Варианты:
a) Режим одиночного узла
b) Работа на одном хосте с отдельными процессами
c) Распределённая работа на множестве узлов
d) Автономный режим без сети
e) Тестовый режим для проверки работоспособности

Вопрос №26
Что такое Heartbeat в HDFS?
Варианты:
a) Протокол голосовой связи
b) Сигнал тревоги при перегрузке сети
c) Сообщение состояния от DataNode к NameNode
d) Событие входа пользователя в систему
e) Электронное письмо с уведомлением

Вопрос №27
Что означает термин Block Report в HDFS?
Варианты:
a) Журнал регистрации пользователей
b) Сообщение о состоянии блоков от DataNode
c) Внутренний протокол авторизации
d) Список загружаемых файлов
e) План задач на ближайший период

Вопрос №28
Что такое линия родства (Lineage) в Apache Spark?
Варианты:
a) Граф зависимости данных и операций
b) История версий программного обеспечения
c) Информационная панель администратора
d) Каталог встроенных библиотек

e) Индекс сортировки данных

Вопрос №29
Что такое ленивая оценка (Lazy Evaluation) в Apache Spark?
Варианты:
a) Выполнение операций только при действии (action)
b) Предварительное кэширование всех данных
c) Параллельное исполнение всех задач
d) Повторное сохранение всех промежуточных результатов
e) Удаление неиспользуемых данных

Вопрос №30
Что обеспечивает отказоустойчивость в распределённых системах?
Варианты:
a) Реализация единой точки отказа
b) Применение принципов бэкапа и репликации
c) Использование отдельного жесткого диска
d) Установка антивирусного ПО
e) Контроль версии файлов

Вопрос №31
Для какой задачи применяется Hadoop?
Варианты:
a) Транзакционные базы данных
b) Потоковая обработка в реальном времени
c) Пакетная обработка больших объёмов данных
d) Виртуализация рабочих станций
e) Микроплатежи и платежи малых сумм

Вопрос №32
Где применяется Apache Spark?
Варианты:
a) Мониторинг сетевого трафика
b) Многопоточные вычисления с быстрой обработкой
c) Онлайн-чаты и мессенджеры
d) Управление серверами DNS
e) Видеоконференции

Вопрос №33
Что характерно для Cassandra?
Варианты:
a) Оптимизация для редких чтений
b) Высокая скорость записи данных
c) Строгость структуры данных
d) Поддержка транзакций ACID
e) Медленность операций записи

Вопрос №34
Какую структуру данных поддерживает MongoDB?
Варианты:
a) Массивы чисел

b) Документы в формате JSON
c) Табличные базы данных
d) Матрицы и векторы
e) Деревья данных

Вопрос №35
Какой критерий важен при выборе между Hadoop и Cassandra?
Варианты:
a) Необходимость масштабируемой обработки крупных партий данных (batch processing)
b) Критическая необходимость высокой скорости произвольного доступа и низкой латентности (low latency) при операциях чтения и записи
c) Требуемая степень согласованности данных (consistency level)
d) Поддержка транзакционности ACID и строгих гарантий целостности данных
e) Предполагаемые сценарии аварийного восстановления и устойчивости к потере данных

Тема 3 «Статистический анализ и визуализация данных»
Вопрос №1
Что такое предварительная обработка данных?
Варианты:
a) Очистка, преобразование и нормализация данных для последующего анализа
b) Построение визуализаций и графиков
c) Сбор данных из открытых источников
d) Создание моделей машинного обучения
e) Импорт данных из внешнего хранилища

Вопрос №2
Какую задачу решает этап очистки данных?
Варианты:
a) Создание индекса для ускорения запросов
b) Устранение пропусков, выбросов и ошибок
c) Автоматизация сбора данных
d) Визуализация данных
e) Генерация синтетических данных

Вопрос №3
Что такое выбросы в данных?
Варианты:
a) Переменные с низким уровнем значимости
b) Значения, резко отличающиеся от общей массы данных
c) Атрибуты, влияющие на конечный результат
d) Источники временных рядов
e) Признаки, введённые вручную

Вопрос №4
Какое средство используется для диагностики выбросов?
Варианты:
a) Диаграмма разброса (Scatter Plot)
b) График автокорреляции (ACF)

c) Коробчатая диаграмма (Box Plot)
d) Круговая диаграмма (Pie Chart)
e) Тепловая карта (Heatmap)

Вопрос №5
Какая операция необходима для приведения данных к общему масштабу?
Варианты:
a) Кодирование категориальных признаков
b) Корреляционный анализ
c) Масштабирование (Scaling)
d) Ансамблевые методы
e) Преобразование Фурье

Вопрос №6
Что такое нормализация данных?
Варианты:
a) Стандартизация данных путем центрирования и масштабирования
b) Создание индексов для ускоренного поиска
c) Преобразование текстовых данных в числовой вид
d) Анализ распределения признаков
e) Визуализация пространственного положения точек

Вопрос №7
Какой метод используется для анализа формы распределения данных?
Варианты:
a) Карта рассеяния (Scatter Matrix)
b) Парный коэффициент корреляции
c) Q-Q Plot (график квантилей)
d) Радиальная диаграмма (Radial Plot)
e) Столбчатая диаграмма (Bar Chart)

Вопрос №8
Что измеряется стандартным отклонением?
Варианты:
a) Центральную тенденцию данных
b) Межквартильный размах данных
c) Изменчивость или разброс данных вокруг среднего
d) Коэффициент асимметрии данных
e) Эффект мультиколлинеарности

Вопрос №9
Какой тест применяется для проверки нормального распределения данных?
Варианты:
a) Тест Шапиро-Уилка
b) Критерий χ^2 (хи-квадрат)
c) t-критерий Стьюдента
d) Критерий Краскела-Уоллиса
e) Метод главных компонент (PCA)

Вопрос №10
Какая мера центрального положения наименее чувствительна к выбросам?

Варианты:
a) Среднее арифметическое
b) Медиана
c) Мода
d) Квантиль
e) Мин/макс

Вопрос №11
Что показывает ковариация между двумя переменными?
Варианты:
a) Линейную связь между переменными
b) Абсолютную величину различий
c) Взаимосвязь временного ряда
d) Нормирующий фактор для средних величин
e) Вероятность взаимного влияния

Вопрос №12
Какой параметр отражает симметрию распределения данных?
Варианты:
a) Асимметрия (Skewness)
b) Коэффициент корреляции
c) Интервал уверенности
d) Среднее геометрическое
e) Гетероскедастичность

Вопрос №13
Что характеризует нормальное распределение данных?
Варианты:
a) Узкий пик и длинные хвосты
b) Смещённость вправо или влево
c) Гауссова форма с симметричностью вокруг среднего
d) Положительная асимметрия
e) Отклонение от средней линии

Вопрос №14
Что такое нулевая гипотеза (H_0)?
Варианты:
a) Гипотеза, принимаемая исследователем изначально
b) Утверждение о наличии/опровержении существования эффекта
c) Заявление, утверждающее существование разницы или связи
d) Альтернативная гипотеза исследования
e) Прогноз, сделанный на основании прошлого опыта

Вопрос №15
Какой тест используется для сравнения средних двух выборок?
Варианты:
a) Тест Уилкоксона
b) t-критерий Стьюдента
c) ANOVA
d) U-критерий Манна-Уитни

e) Тест Крускала-Уоллиса

Вопрос №16
Какой критерий применяется для оценки однородности дисперсий?
Варианты:
a) Ljung-Box тест
b) Байесовский критерий
c) Тест Бартлетта
d) Тест Шарпа-Линднера
e) ROC-AUC

Вопрос №17
Какой метод используется для коррекции множественных сравнений?
Варианты:
a) Поправка Бонферрони
b) Максимизация правдоподобия
c) Метод ближайшего соседа
d) Метод главных компонент
e) Обратное распространение ошибки

Вопрос №18
Что описывает матрица корреляций?
Варианты:
a) Влияние каждой переменной на остальные
b) Направление и силу линейной связи между переменными
c) Интервалы возможных значений
d) Структуру временных рядов
e) Границы доверительного интервала

Вопрос №19
Что такое бутстреп-метод?
Варианты:
a) Техника машинного обучения
b) Процедура выравнивания данных
c) Итерационное повторное взятие выборок для оценки точности
d) Оптимизация гиперпараметров
e) Нормализация данных

Вопрос №20
Какой инструмент предназначен для автоматической разведки данных (EDA)?
Варианты:
a) Scikit-Learn
b) Sweetviz
c) TensorFlow
d) PyTorch
e) Seaborn

Вопрос №21
Что такое гистограмма (histogram)?
Варианты:
a) Вид графика, демонстрирующий частоту появления значений одной

переменной

b) График, показывающий изменение переменной во времени

c) Диаграмма, представляющая долю каждой категории в общем количестве

d) Карта, иллюстрирующая территориальное распределение

e) Диаграмма, отражающая зависимость между тремя переменными

Вопрос №22

Что представляют собой круговые диаграммы (pie chart)?

Варианты:

a) Наглядное изображение процентного соотношения частей целого

b) Демонстрация динамики показателя во времени

c) Представление распределения плотности данных

d) Иллюстрация временной последовательности событий

e) График корреляции между двумя переменными

Вопрос №23

Что такое древовидная карта (tree map)?

Варианты:

a) Визуализация иерархической структуры данных с отображением размеров объектов

b) Диаграмма рассеяния для демонстрации взаимосвязи двух переменных

c) Картографический аналог тепловых карт

d) Вид графика для анализа временных рядов

e) Представление многомерных данных в виде сетки

Вопрос №24

Какой тип графика используется для демонстрации зависимости между двумя переменными?

Варианты:

a) График рассеяния (scatter plot)

b) Линейный график (line chart)

c) Гистограмма (histogram)

d) Древовидная карта (treemap)

e) Круговая диаграмма (pie chart)

Вопрос №25

Что показывает тепловая карта (heatmap)?

Варианты:

a) Изменение показателя во времени

b) Долю каждой категории в общем объёме

c) Интенсивность взаимоотношений между переменными

d) Иерархическую структуру данных

e) Пространственное распределение явления

Вопрос №26

Какой принцип лежит в основе хорошей визуализации данных?

Варианты:

a) Максимальная насыщенность цветами

b) Простота и ясность подачи информации

c) Преимущественно использование круговых диаграмм

d) Нарушение пропорций и масштабов

e) Доминирование интерактивных элементов

Вопрос №27

Какой график лучше всего демонстрирует динамику изменения параметра во времени?

Варианты:

a) Линейный график (line chart)
b) Гистограмма (histogram)
c) Тепловая карта (heatmap)
d) Диаграмма рассеяния (scatter plot)
e) Древовидная карта (treemap)

Вопрос №28

Что такое диаграмма рассеяния (scatter plot)?

Варианты:

a) Вид графика, показывающий взаимосвязь между двумя переменными
b) Форма круговой диаграммы для иллюстрации долей
c) Представление иерархической структуры данных
d) График для анализа сезонных колебаний
e) Наглядное отражение частоты событий

Вопрос №29

Что является недостатком использования круговых диаграмм (pie chart)?

Варианты:

a) Удобство восприятия данных
b) Хорошая наглядность при большом количестве категорий
c) Трудность понимания при наличии множества мелких сегментов
d) Эффективность отображения небольших различий
e) Возможность выделить важную категорию цветом

Вопрос №30

Какой инструмент визуализации полезен для представления сложной иерархической структуры данных?

Варианты:

a) Древовидная карта (treemap)
b) Линейный график (line chart)
c) Диаграмма рассеяния (scatter plot)
d) Гистограмма (histogram)
e) Тепловая карта (heatmap)

Вопрос №31

Что из нижеперечисленного верно для инструмента Tableau?

Варианты:

a) Используется только для анализа временных рядов
b) Подходит для создания интерактивных дашбордов
c) Применяется исключительно для финансового анализа
d) Является бесплатным инструментом для коммерческого использования
e) Поддерживает только текстовые форматы данных

Вопрос №32

Какой инструмент широко используется для анализа и визуализации больших объёмов данных в бизнесе?

Варианты:

a) Paint

b) Tableau

c) Adobe Illustrator

d) Gephi

e) Unity

Вопрос №33

Какой инструмент идеально подходит для интеграции с продуктами Microsoft и удобен для построения интерактивных отчетов?

Варианты:

a) Power BI

b) Photoshop

c) Maya

d) Blender

e) GIMP

Вопрос №34

Что представляет собой библиотека Matplotlib?

Варианты:

a) Бесплатный векторный графический редактор

b) Платформу для совместного редактирования текста

c) Библиотеку Python для построения графиков

d) Систему для создания анимаций

e) Инструмент для дизайна веб-интерфейсов

Вопрос №35

Какой инструмент из перечисленных позволяет создавать красивые и информативные графики с минимальной настройкой?

Варианты:

a) Visio

b) PowerPoint

c) Seaborn

d) CorelDRAW

e) Excel

Вопрос №36

Какой инструмент поддерживает широкую палитру цветовых схем и стильный внешний вид графиков?

Варианты:

a) Matplotlib

b) Excel

c) Power BI

d) Seaborn

e) Word

Вопрос №37

Какой инструмент визуализации обладает широким спектром интерактивных элементов и удобен для анализа в браузере?

Варианты:
a) Photoshop
b) Paint
c) Plotly
d) Excel
e) PowerPoint

Вопрос №38

Какой инструмент лучше всего подойдёт для визуализации данных на основе стандартных аналитических отчётов?

Варианты:
a) Maya
b) Matplotlib
c) Power BI
d) Cinema 4D
e) Lightroom

Вопрос №39

Какой инструмент Python является низкоуровневым и допускает полную кастомизацию графиков?

Варианты:
a) Seaborn
b) Plotly
c) Matplotlib
d) Bokeh
e) Dash

Вопрос №40

Какой инструмент рекомендован для новичков в визуализации данных, желающих быстро освоить создание графиков в Python?

Варианты:
a) Mayavi
b) Pandas
c) Seaborn
d) NumPy
e) PyGame

Вопрос №41

Что такое панель мониторинга (dashboard)?

Варианты:
a) Средство визуализации и оперативного анализа данных
b) Документ с зафиксированной историей изменений
c) Программный модуль для интеграции баз данных
d) Техническое руководство для регламентных работ
e) Шаблон для отчетности финансовой деятельности

Вопрос №42

Какие критерии определяют эффективность панели мониторинга?

Варианты:
a) Простота структуры и удобная навигация

b) Яркие и контрастные цвета фона
c) Применение большого количества графиков
d) Частое обновление данных
e) Наличие списка ссылок на источники данных

Вопрос №43
Что обязательно должно присутствовать на эффективной панели мониторинга?
Варианты:
a) Разнообразные фильтры и кнопки экспорта
b) Визуализации ключевых показателей (KPI)
c) Тексты инструкций и пояснений
d) Исторические справки
e) Иллюстрации

Вопрос №44
Что является важнейшим элементом интерактивной панели мониторинга?
Варианты:
a) Возможность постоянного обновления данных
b) Присутствие контактной информации
c) Настройка фиксированного режима отображения
d) Применение логотипа компании
e) Составление плана развития

Вопрос №45
Какой элемент способствует быстрому погружению в детали на панели мониторинга?
Варианты:
a) Панель переключения режимов просмотра
b) Функция детального анализа ("drill-down")
c) Календарь с отметками
d) Формы обратной связи для предложений пользователей
e) Список основных выводов

Вопрос №46
Какую информацию предоставляет Google Analytics в панели мониторинга?
Варианты:
a) Результаты внутренней бухгалтерии компании
b) Посещаемость сайта и поведение пользователей
c) Список действующих контрактов и соглашений
d) Статистику закупок сырья и материалов
e) Информацию о кадровых изменениях в коллективе

Вопрос №47
Какое важное отличие панели мониторинга от обычного отчета?
Варианты:
a) Панель мониторига предоставляется только начальникам отделов
b) Панель мониторинга всегда статична и не изменяется
c) Панель мониторинга интерактивна и обновляется автоматически
d) Панель мониторинга предназначена только для внутреннего пользования
e) Панель мониторинга формируется вручную специалистами

Вопрос №48

Какой тип панели мониторинга используется для фиксации краткосрочных изменений данных?

Варианты:

a) Аналитический

b) Стратегический

c) Операционный

d) Долгосрочный

e) Справочный

Вопрос №49

Что необходимо учитывать при проектировании макета панели мониторинга?

Варианты:

a) Соответствие цветовой гаммы фирменному стилю компании

b) Оптимальное расположение блоков и простота восприятия данных

c) Сезонные предпочтения пользователей в оформлении

d) Целевую аудиторию и её потребности в информации

e) Средний возраст сотрудников, пользующихся панелью

Вопрос №50

Какой этап в методологии построения панели мониторинга предшествует созданию макета?

Варианты:

a) Оформление акта сдачи-приёма готового проекта

b) Рассылка приглашений на совещание по дизайну

c) Подготовка подробного пресс-релиза для прессы

d) Определение целевых аудиторий и потребностей

e) Подписание договора с поставщиками услуг

Тема 4 «Машинное обучение и методы классификации и кластеризации»

Вопрос №1

Что такое формальное определение обучения по Т. Митчеллу?

a) Процесс улучшения показателей программы при выполнении конкретной задачи без внешнего вмешательства.

b) Улучшение способности компьютера решать задачи на основе собственной интуиции.

c) Автоматическое улучшение программой своей производительности на задаче T, основываясь на опыте E и мере оценки эффективности P.

d) Программирование машины для решения узко специализированных задач.

e) Постепенный рост производительности системы благодаря увеличению объема памяти.

Вопрос №2

Какова основная цель обучения с учителем?

a) Минимизация ошибки предсказания на размеченных данных.

b) Поиск скрытой структуры в данных.

c) Оптимизация вычислительных затрат на обработку данных.

d) Максимизация точности модели на тестовых данных.

e) Уменьшение шума в сигналах датчиков.

Вопрос №3

Чем отличается обучение без учителя от обучения с учителем?

a) Использование исключительно непрерывных переменных.

b) Отсутствие предварительных меток на обучающих данных.

c) Требование больших объемов размеченных данных.

d) Ограничение числа возможных решений.

e) Невозможность применять глубокое обучение.

Вопрос №4

Какой тип проблемы решается методами обучения без учителя?

a) Задача классификации образов.

b) Регрессия временных рядов.

c) Выделение общих паттернов в данных.

d) Определение целевой аудитории продукта.

e) Прогнозирование курса акций.

Вопрос №5

Что характеризует полуправленное обучение?

a) Полностью автоматизированный процесс обучения.

b) Только вручную промаркированные данные.

c) Небольшой объем размеченных данных плюс большой объем неподготовленных данных.

d) Полная зависимость от эксперта.

e) Исключение возможности ошибок в процессе обучения.

Вопрос №6

Какие преимущества имеет обучение с учителем перед другими видами обучения?

a) Высокая эффективность даже при небольшом количестве размеченных данных.

b) Возможность автоматического повышения качества данных.

c) Высокое качество предсказаний при достаточном количестве размеченных данных.

d) Легкость интеграции в распределённые системы.

e) Устойчивость к шумовым выбросам.

Вопрос №7

Каково основное назначение методов обучения без учителя?

a) Предсказание будущих событий.

b) Получение структурированных выводов из сложных данных.

c) Решение задач бинарной классификации.

d) Проверка гипотез статистическими методами.

e) Оптимизация финансовых инвестиций.

Вопрос №8

Что означает термин "кластеризация"?

a) Подгонка гиперплоскостей для наилучшего соответствия данных.

b) Разбиение множества элементов на подгруппы, обладающие общими характеристиками.

c) Преобразование векторных пространств в скалярные величины.

d) Коррекция предикторов в уравнении регрессии.

e) Очистка датасета от дубликатов записей.

Вопрос №9

Каковы недостатки обучения без учителя?

a) Сложность реализации на графических процессорах.

b) Низкая скорость работы на небольших наборах данных.

c) Трудность интерпретации найденных группировок.

d) Необходимость большого количества предварительно обработанной информации.

e) Склонность к переобучению на малочисленные шаблоны.

Вопрос №10

В чём преимущество использования Autoencoders в полуправленном обучении?

a) Способность обнаруживать точные вероятности распределения классов.

b) Автоматический отбор наиболее важных признаков.

c) Эффективность реконструкции данных и выделения внутренних закономерностей.

d) Быстрое восстановление слабоструктурированных данных.

e) Увеличение стабильности модели при малых изменениях условий среды.

Вопрос №11

Что описывает логистическая регрессия?

a) Вероятность принадлежности элемента к определенному классу.

b) Распределение гауссианов в пространстве признаков.

c) Градиентный спуск для оптимизации потерь.

d) Функциональную связь между признаками и целевым признаком.

e) Временную последовательность событий.

Вопрос №12

Почему используется кросс-энтропийный показатель потери (Cross Entropy Loss)?

a) Для упрощения процесса нормализации признаков.

b) Чтобы повысить устойчивость модели к зашумленным данным.

c) Для измерения различия между двумя вероятностными распределениями.

d) Из-за низкой чувствительности к небольшим изменениям веса.

e) Для ускорения вычислений в параллельных системах.

Вопрос №13

В каком сценарии применяется K-means алгоритм?

a) Когда важно сохранить порядок следования данных.

b) Если данные требуют точной временной привязки.

c) Когда требуется разделить данные на однородные группы.

d) Если важна интерпретация каждого отдельного примера.

e) Если необходимы строгие ограничения на количество используемых признаков.

Вопрос №14

Какая главная проблема возникает при применении метода Principal Component Analysis (PCA)?

a) Потеря части оригинальной информации при уменьшении размерности.

b) Повышенная сложность в обработке бинарных признаков.

c) Несоответствие результатов требованиям бизнеса.

d) Риск появления циклических зависимостей.

e) Неустойчивость к малым отклонениям в значениях признаков.

Вопрос №15

Как называется методика, используемая для изоляции необычных наблюдений среди данных?

a) Cross Validation.
b) Anomaly Detection.
c) Boosting.
d) Ensemble Learning.
e) Overfitting Prevention.

Вопрос №16

Какова ключевая особенность полуправленного обучения?

a) Использование лишь подмножества размеченных данных вместе с большим объемом неподготовленных данных.
b) Возможность быстрого изменения решаемых задач.
c) Максимальная адаптация модели к разным областям применения.
d) Применение только сетевых архитектур глубокого обучения.
e) Непосредственный контроль над процессом подбора функций активации.

Вопрос №17

Какой из перечисленных методов чаще всего применяют для уменьшения размерности данных?

a) Gradient Descent.
b) Principal Component Analysis (PCA).
c) Logistic Regression.
d) Support Vector Machines (SVM).
e) Backpropagation.

Вопрос №18

Какая мера помогает оценить степень разброса данных вокруг среднего значения?

a) Стандартное отклонение.
b) Среднее абсолютное отклонение.
c) Коэффициент корреляции Пирсона.
d) Среднеквадратичное отклонение.
e) Интервал доверия.

Вопрос №19

Какова главная особенность Self-Trainings по сравнению с классическим обучением с учителем?

a) Самостоятельное расширение размеченного набора данных путём предсказания меток.
b) Необходимость постоянного контроля специалиста над процессом обучения.
c) Возможность адаптации только к непрерывным признакам.
d) Использование большего количества серверов для параллельной обработки.
e) Невозможность повторяемости экспериментов.

Вопрос №20

Как метод Isolation Forest определяет аномалию?

a) Путём случайного построения дерева, которое выделяет необычные объекты раньше обычных.

b) Используя градиенты для оценки важности признаков.

c) Применяя технику ансамбля классификаторов для подтверждения аномальности.

d) Оценивая влияние конкретного признака на общую статистику.

e) Рассчитывая коэффициент доверительной границы.

Вопрос №21

Что такое логистическая регрессия?

a) Линейная модель для прогнозирования количественных признаков.

b) Ансамблевая техника для объединения разных моделей.

c) Метод для определения связи между несколькими категориями.

d) Алгоритм для решения задач бинарной классификации.

e) Техника для группировки сходных объектов.

Вопрос №22

Какую меру используют для оценки качества модели логистической регрессии?

a) Mean squared error (MSE).

b) Logarithmic loss (log loss / cross entropy loss).

c) Accuracy only for balanced datasets.

d) Information gain.

e) Precision-recall curve.

Вопрос №23

Что понимается под деревьями решений в контексте машинного обучения?

a) Методы факторного анализа.

b) Последовательное деление выборки на подгруппы на основе признаков.

c) Совокупность классических статистических тестов.

d) Классический байесовский подход.

e) Специализированные искусственные нейронные сети.

Вопрос №24

Какой критерий используют для определения степени увеличения информативности при построении дерева решений?

a) F-статистику.

b) Information gain (увеличение информации).

c) Критерий Хи-квадрат.

d) ANOVA-тест.

e) Метод наименьших квадратов.

Вопрос №25

Что такое случайный лес (random forest)?

a) Один сложный древовидный классификатор.

b) Набор маленьких деревьев решений, созданных случайным образом.

c) Группа искусственных нейронов.

d) Эволюционный алгоритм генетического программирования.

e) Детектор краёв в изображениях.

Вопрос №26

В чем заключается основной принцип градиентного бустинга?

a) Объединение множества базовых моделей параллельно.

b) Добавление нового слоя нейронов в глубокую нейронную сеть.

c) Постройка деревьев решений независимо друг от друга.

d) Последовательное создание моделей, каждая из которых уменьшает ошибки предыдущих.

e) Выбор лучших признаков на основе анализа главных компонентов.

Вопрос №27

Какой эффект достигается применением случайного леса по сравнению с одиночным деревом решений?

a) Рост скорости обучения.

b) Снижение риска переобучения.

c) Возрастание сложности настройки модели.

d) Увеличение зависимости от количества признаков.

e) Ухудшение общей производительности.

Вопрос №28

Что означают γ и $h_m(x)$ в формуле градиентного бустинга?

a) Количество деревьев и новый слой нейронов соответственно.

b) Скорость обучения и ошибка прошлой модели соответственно.

c) Новый уровень в дереве и старый набор признаков соответственно.

d) Вес поправки и новое базовое дерево соответственно.

e) Значения стандартной ошибки и градиента функции потерь соответственно.

Вопрос №29

Какую проблему решает регуляризация в градиентном бустинге?

a) Проблему нехватки признаков.

b) Переобучение модели.

c) Недостаточную гибкость модели.

d) Слишком быстрое снижение ошибок.

e) Большую чувствительность к отсутствующим данным.

Вопрос №30

Какой алгоритм лучше выбрать, если важен простой и прозрачный вывод?

a) Случайный лес.

b) Градиентный бустинг.

c) Деревья решений.

d) Глубокое обучение.

e) Байесовская статистика.

Вопрос №31

Что такое кластерный анализ?

a) Это процедура автоматической сегментации данных на группы (кластеры) на основе сходства объектов.

b) Способ классификации объектов на заранее известные категории.

c) Тип обучения с учителем, где каждый объект уже обладает меткой.

d) Метод визуального представления данных в двумерном пространстве.

e) Инструмент для снижения размерности данных.

Вопрос №32

Какова основная цель алгоритма k-means?

a) Формирование произвольных фигур кластеров.
b) Нахождение глобального минимума функции потерь.
c) Разделение данных на заранее известное количество кластеров.
d) Автоопределение оптимальной архитектуры нейронной сети.
e) Распространение сигнала по нейронам многослойной сети.

Вопрос №33
Какая формула отражает основную идею k-means?

a) $J(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$
b) $\nabla_w L(w) = -\frac{1}{n} \sum_{i=1}^{n} y_i x_i + \lambda w$
c) $H(Y|X) = H(X) + H(Y) - I(X;Y)$
d) $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
e) $f(x) = w^T \phi(x) + \epsilon$

Вопрос №34
Как называется этап присвоения объектов к ближайшим центрам кластеров в k-means?
a) Стадия инициализации.
b) Шаг агрегирования.
c) Фаза перераспределения.
d) Итерация смещения.
e) Процедурный этап аппроксимации.

Вопрос №35
Что обозначает символ μ_i в формуле суммы квадратов расстояний k-means?
a) Центр масс соответствующего кластера C_i.
b) Масштабирующий множитель функции активации.
c) Дисперсия внутри выбранного кластера.
d) Величина штрафа за переобучение.
e) Функция перехода состояний.

Вопрос №36
Какой недостаток присущ методу k-means?
a) Чрезвычайно долгая работа на больших наборах данных.
b) Высокая требовательность к объему оперативной памяти.
c) Обязательное задание точного количества кластеров заранее.
d) Крайняя восприимчивость к пропускам данных.
e) Полная неспособность выявлять шум.

Вопрос №37
Что такое иерархическое кластирование?
a) Алгоритм обучения с учителем, использующий метки классов.
b) Процедура кластеризации, создающая дерево включенности кластеров.
c) Специальный тип логистической регрессии.
d) Особенность метода главных компонент.
e) Способ выравнивания частот встречаемости признаков.

Вопрос №38
Что представляет собой single-linkage в иерархическом кластировании?

a) Стратегия объединения кластеров по самому большому расстоянию между любыми двумя объектами.

b) Подход слияния на основе средней дистанции между объектами.

c) Метод соединения по минимальной дистанции между парой ближайших объектов.

d) Агрегативный метод выбора размера окна скользящего среднего.

e) Точность предсказательной силы на проверке на новых данных.

Вопрос №39

Какой подход называют divisive в иерархическом кластировании?

a) Параллельное выполнение операций.

b) Принцип объединения близких объектов в один кластер.

c) Постепенное деление единого целого на отдельные подгруппы.

d) Перебор всех комбинаций признаков.

e) Удаление всех выделившихся аномалий.

Вопрос №40

В чем особенность DBSCAN по сравнению с k-means?

a) DBSCAN способен формировать кластеры различной формы и размеров.

b) DBSCAN обязательно требует задания количества кластеров.

c) DBSCAN чрезвычайно быстр на крупных наборах данных.

d) DBSCAN идеально подходит для категориальных данных.

e) DBSCAN гарантирует абсолютно стабильную работу вне зависимости от начальных условий.

Вопрос №41

Что означает термин "ядро" (core point) в DBSCAN?

a) Объекты, имеющие достаточно соседей в пределах заданного радиуса.

b) Внутренняя часть любого отдельно взятого кластера.

c) Главная ось симметрии фигуры кластера.

d) Границы раздела между кластерами.

e) Нарушение непрерывности поверхности при поиске ближайших соседей.

Вопрос №42

Какой параметр определяет минимальный размер плотного скопления объектов в DBSCAN?

a) Radius (r).

b) Min_samples.

c) Alpha.

d) Beta.

e) Lambda.

Вопрос №43

Что является основным недостатком DBSCAN?

a) Сложность подбора ключевых параметров (eps и min_samples).

b) Медленность работы на больших размерах данных.

c) Неточное воспроизведение геометрических форм кластеров.

d) Низкая чувствительность к изменению масштаба признаков.

e) Постоянное образование огромного количества мелких кластеров.

Вопрос №44

Какой метод рекомендуется использовать, если имеется небольшое количество данных с хорошо выраженным числом кластеров?

a) K-means.
b) Иерархическое кластирование.
c) DBSCAN.
d) Gaussian mixture model.
e) Deep learning.

Вопрос №45

Какова роль функции потери в k-means?
a) Она контролирует глубину дерева решений.
b) Используется для оптимизации позиционирования центров кластеров.
c) Определяет размеры слоев нейронной сети.
d) Показатель успешности обучения в логистической регрессии.
e) Мерило качества сверточных фильтров.

Вопрос №46

Какой тип кластеров предпочтительнее выбирать при работе с неоднородными по форме группами данных?
a) Плотные компактные кругообразные кластеры.
b) Произвольные выпуклые кластеры.
c) Связанные по плотности области.
d) Округлые кластеры с плавными границами.
e) Ячейковые кластеры сетки.

Вопрос №47

Что такое border points в DBSCAN?
a) Важнейшие опорные точки кластера.
b) Исключительные внешние элементы, удаленные от основного массива.
c) Окружающие пограничные точки ядра кластера.
d) Ключевые вершины графа связности.
e) Контрольные точки для проверки надежности.

Вопрос №48

Что означает понятие noise в контексте DBSCAN?
a) Случайные флуктуации уровня энергии сигналов.
b) Отдельные элементы, не относящиеся ни к какому кластеру.
c) Отклонения измерений от реальных значений.
d) Количественно определенный минимум разброса данных.
e) Исходящие ребра ориентационного графа.

Вопрос №49

В каком случае целесообразно использовать метод k-means?
a) При желании исследовать внутреннюю структуру разнородных данных.
b) Если известно приблизительное количество ожидаемых кластеров.
c) Когда необходима детализация форм границ кластеров.
d) Если предстоит анализировать графовые структуры.
e) При работе с категоричными данными.

Вопрос №50

Какая рекомендация дана для правильной подготовки данных перед

использованием k-means?

a) Всегда нормализовать признаки.

b) Игнорировать наличие выбросов.

c) Использовать только целые числа.

d) Никогда не удалять пропущенные значения.

e) Применять только одно правило разбиения на группы.

Тема 5 «Глубокое обучение и нейронные сети»

Вопрос №1

Что такое искусственная нейронная сеть?

a) Система из простых элементов обработки информации, аналогичных биологическим нейронам.

b) Сборщик данных для предварительной обработки информации.

c) Одноклеточный организм, способный размножаться.

d) Средство визуализации многомерных данных.

e) Модель передачи сообщений между пользователями.

Вопрос №2

Какие основные компоненты входят в состав искусственного нейрона?

a) Входные сигналы, тело нейрона, сумма, функция активации.

b) Матрицы, операции умножения и свертки.

c) Аккумулятор, аккумуляторные батареи и зарядные устройства.

d) Антенна, усилители мощности и фильтры.

e) Переменные состояния, внутренние регистры и буферы.

Вопрос №3

Какое предназначение имеет функция активации в искусственном нейроне?

a) Хранение постоянных величин.

b) Приведение выходного сигнала в нужный диапазон значений.

c) Генерация уникальных идентификационных номеров.

d) Управление памятью компьютера.

e) Обеспечение интерфейса ввода-вывода.

Вопрос №4

Какой тип функции активации обеспечивает преобразование значений в диапазоне [0,1]?

a) Сигмоидная функция.

b) Гиперболический тангенс.

c) Rectified Linear Unit (ReLU).

d) Степенная функция.

e) Exponential Linear Unit (ELU).

Вопрос №5

Что представляет собой входной слой в структуре нейронной сети?

a) Первый слой, принимающий начальные данные.

b) Последний слой, выдающий итоговый результат.

c) Промежуточный слой, обрабатывающий данные.

d) Вспомогательная память для хранения промежуточных расчетов.

e) Внешнюю оболочку для защиты внутренней структуры.

Вопрос №6

Как называются внутренние слои нейронной сети, находящиеся между входным и выходным слоями?

a) Видимые слои.

b) Открытые слои.

c) Активные слои.

d) Скрытые слои.

e) Дополнительные слои.

Вопрос №7

Какой тип нейронной сети применяется преимущественно для обработки изображений?

a) Рекуррентные нейронные сети (RNN).

b) Генеративно-состязательные сети (GAN).

c) Сверточные нейронные сети (CNN).

d) Однослойные перцептроны.

e) Многослойные персептроны (MLP).

Вопрос №8

Какой механизм лежит в основе обучения нейронных сетей методом обратного распространения ошибки (backpropagation)?

a) Прямой расчет средних значений.

b) Редукция данных с помощью фильтрации.

c) Пропуск сигнала вперед и последующее обновление весов в обратном направлении.

d) Локализация пиковых значений.

e) Резервное копирование и восстановление данных.

Вопрос №9

Какая проблема связана с глубокими нейронными сетями при использовании метода обратного распространения ошибки?

a) Проблема исчезновения градиентов.

b) Упрощение архитектуры сети.

c) Высокоэффективная распараллеливаемость вычислений.

d) Безошибочные прогнозы.

e) Абсолютная независимость результата от входных данных.

Вопрос №10

Как называется явление, при котором нейронная сеть становится слишком сложной и теряет способность правильно обобщать данные?

a) Недообучение.

b) Переобучение.

c) Полноценное обучение.

d) Частичное обучение.

e) Адаптационное обучение.

Вопрос №11

Что представляет собой выходное значение простого искусственного нейрона?

a) Результат суммирования входных сигналов.

b) Произведение входных сигналов и весов.

c) Выход после прохождения через функцию активации.

d) Абсолютное значение входных сигналов.

e) Константу, равную единице.

Вопрос №12

В формуле выхода нейрона '*y*' это?

a) Вес связи нейрона.

b) Смещение нейрона.

c) Суммарный вход нейрона.

d) Выходной сигнал нейрона.

e) Индекс номера нейрона.

Вопрос №13

Какая функция активации возвращает положительное значение, если аргумент положительный, иначе возвращает нулевое значение?

a) Сигмоида.

b) Hyperbolic tangent (tanh).

c) Rectified Linear Unit (ReLU).

d) Arctangent.

e) Softmax.

Вопрос №14

Какую задачу решают рекуррентные нейронные сети (RNN)?

a) Представление статических изображений.

b) Обработка последовательных данных (тексты, звук, временные ряды).

c) Трансформация сигналов в частоты Фурье.

d) Решение дифференциальных уравнений второго порядка.

e) Генерация музыкальных композиций в стиле классической музыки.

Вопрос №15

Какая операция проводится в первом этапе алгоритма обратного распространения ошибки?

a) Выполнение прямого прохода (forward pass).

b) Расчет обратной передачи сигнала.

c) Изменение весов на противоположные знаки.

d) Нормализация входных данных.

e) Применение функции активации ко всему слою одновременно.

Вопрос №16

Для какого типа данных чаще всего используются свёрточные нейронные сети (CNN)?

a) Двумерные данные, такие как изображения.

b) Последовательные данные, такие как речь и текст.

c) Табличные данные с множеством столбцов.

d) Временные ряды с нерегулярными промежутками.

e) Биометрические данные, такие как отпечатки пальцев.

Вопрос №17

Что такое фильтр (kernel) в свёрточных нейронных сетях?

a) Малый участок изображения, на котором производится свёртка.

b) Метод понижения размерности карт признаков.

c) Узкий прямоугольник, охватывающий границу объекта.

d) Специальный алгоритм сортировки пикселей.

e) Многоразрядный регистр для временного хранения информации.

Вопрос №18

Что означает процесс pooling в свёрточных нейронных сетях?

a) Уменьшение разрешения карты признаков путем выбора максимального или среднего значения.

b) Преобразование цветов RGB в оттенки серого.

c) Сохранение первоначальной топологии входных данных.

d) Запись результатов работы в отдельный файл журнала.

e) Сознательный выбор нежелательных областей изображения.

Вопрос №19

Что отличает рекуррентные нейронные сети (RNN) от стандартных нейронных сетей?

a) Наличие циклических связей, позволяющих хранить информацию о предыдущем состоянии.

b) Отсутствие механизмов обработки длинных последовательностей.

c) Очень высокие требования к аппаратным ресурсам.

d) Аппаратная поддержка только GPU-процессоров.

e) Возможности только для классификации.

Вопрос №20

Какая проблема характерна для классических рекуррентных нейронных сетей (RNN)?

a) Они работают только с одномерными данными.

b) Затруднения с долгосрочными зависимостями из-за исчезающего градиента.

c) Работа исключительно с графиковыми процессорами.

d) Использование чрезмерно большого количества памяти.

e) Нельзя объединить с другими типами нейронных сетей.

Вопрос №21

Что такое ячейка памяти (memory cell) в архитектуре LSTM?

a) Механизм, контролирующий сохранение важной информации в течение долгого времени.

b) Альтернативный метод замены рекуррентных связей.

c) Устройство хранения в классическом компьютере.

d) Некоторое подобие файла cookie в браузере.

e) Другое название таймера ожидания.

Вопрос №22

Что выполняет забывающий вентиль (forget gate) в ячейке LSTM?

a) Определяет, какую информацию оставить в ячейке памяти.

b) Включает режим резервного копирования данных.

c) Блокирует доступ к внешним источникам данных.

d) Формирует новые обучающиеся экземпляры.

e) Устанавливает периодичность активности.

Вопрос №23

Что такое внимание (Attention Mechanism) в трансформерах?

a) Способность сети концентрироваться на определенных фрагментах входных данных.

b) Процедура шифрования конфиденциальных данных.
c) Метод переключения режимов активации.
d) Средство устранения неполадок в программе.
e) Функция автоподстановки в текстовом редакторе.

Вопрос №24
Зачем нужен Multi-head Attention в трансформерах?
a) Позволяет одновременно рассматривать разные аспекты данных.
b) Ускоряет процесс конвертации файлов формата PDF.
c) Осуществляет перенос данных между серверами.
d) Помогает записывать журналы регистрации.
e) Создает дополнительное пространство для резервного хранилища.

Вопрос №25
Что делают механизмы внимания (Attention Mechanisms) в трансформерах?
a) Помогают сконцентрироваться на значимых областях входных данных.
b) Производят сжатие данных с потерями.
c) Проводят калибровку сенсоров камеры.
d) Исправляют орфографические ошибки в тексте.
e) Копируют файлы с одного диска на другой.

Вопрос №26
Какая библиотека чаще всего используется для создания свёрточных нейронных сетей (CNN)?
a) TensorFlow/Keras.
b) JavaScript Canvas API.
c) SQL Server Management Studio.
d) Photoshop CC.
e) Excel VBA макросы.

Вопрос №27
Какая функция активации широко применяется в свёрточных нейронных сетях (CNN)?
a) ReLU (Rectified Linear Unit).
b) ZigZag.
c) Cosine similarity.
d) Geometric mean.
e) Weighted sum.

Вопрос №28
В чём преимущество трансформеров перед традиционными рекуррентными нейронными сетями (RNN)?
a) Могут параллельно обрабатывать все входные данные, увеличивая скорость обучения.
b) Поддерживают только табличные данные.
c) Работают только с аудиоданными.
d) Реализуют полный отказ от свёрток.
e) Используют гибридные квантовые процессоры.

Вопрос №29

Какая задача является одной из ключевых целей использования рекуррентных нейронных сетей (RNN)?

a) Обработка последовательных данных, таких как текст и речь.

b) Обнаружение транспортных маршрутов на картах.

c) Каталогизация музейных экспонатов.

d) Управление промышленными роботами.

e) Определение возраста пользователей по фотографиям.

Вопрос №30

Что является главной особенностью архитектуры LSTM?

a) Возможность запоминать информацию на длительный срок.

b) Особый способ преобразования спектрального состава звука.

c) Сворачивание многомерных матриц в одномерные строки.

d) Конвертирование изображений в чёрно-белый формат.

e) Создание иллюзий движения на неподвижных картинках.

Вопрос №31

Что такое свёрточный слой в CNN?

a) Производит операции свёртки с фильтром, выделяя специфические характеристики изображения.

b) Накладывает шаблон с равномерным цветом поверх картинки.

c) Перекрывает картину дополнительным слоем фона.

d) Увеличивает разрешение изображения.

e) Смешивает цвета изображения с целью искажения деталей.

Вопрос №32

Что значит операция pooling в CNN?

a) Уменьшает размерность карты признаков путём выбора максимумов или средних значений.

b) Добавляет шумы к изображению для усиления признаков.

c) Растягивает изображение, делая его шире или длиннее.

d) Меняет цветовую палитру изображения.

e) Заменяет фрагменты изображения рандомными значениями.

Вопрос №33

Какая функция активации обычно используется в CNN?

a) ReLU (Rectified Linear Unit).

b) Синусоидальная функция.

c) Кубическая функция.

d) Экспоненциальная функция.

e) Периодическая функция.

Вопрос №34

Что такое векторизация слов (word embedding)?

a) Представление слов в виде низкоразмерных векторов, сохраняющих смысловую близость.

b) Замена букв на цифры.

c) Удаление стоп-слов из текста.

d) Автоматическое написание заголовков статей.

e) Создание списков синонимов.

Вопрос №35

Что такое рекуррентная нейронная сеть (RNN)?

a) Нейронная сеть, учитывающая предыдущее состояние при обработке текущих данных.

b) Пространственная сетка для обработки изображений.

c) Статистический метод проверки гипотез.

d) Метод сжатия мультимедийных файлов.

e) Процесс сокращения длины строк.

Вопрос №36

Какая проблема возникает в стандартных RNN?

a) Vanishing gradient problem (проблема исчезающих градиентов).

b) Постоянная потребность в дополнительном оборудовании.

c) Низкая скорость выполнения операций.

d) Большая энергоёмкость.

e) Недостаток гибкости в настройке параметров.

Вопрос №37

Что такое Long Short-Term Memory (LSTM)?

a) Улучшенный тип RNN, предназначенный для решения проблемы долговременной зависимости.

b) Модифицированное устройство жестких дисков.

c) Новая технология беспроводной связи.

d) Нейронная сеть, работающая исключительно с текстом.

e) Архитектура для распознавания звуков.

Вопрос №38

Что такое трансформер (Transformer)?

a) Архитектура нейронной сети, созданная для эффективного анализа текстовых данных с механизмом внимания.

b) Электрический прибор для преобразования напряжения.

c) Автомобиль с искусственным интеллектом.

d) Технология виртуальной реальности.

e) Робот-трансформер из мультфильма.

Вопрос №39

Что такое временная серия (Time series)?

a) Ряд данных, упорядоченных во времени, характеризующих развитие процесса.

b) Серия фильмов или сериалов.

c) Текущие цены товаров на рынке.

d) Список спортивных рекордов.

e) Список покупок продуктов питания.

Вопрос №40

Какие нейронные сети эффективны для анализа временных рядов?

a) Рекуррентные нейронные сети (RNN).

b) Спектральные нейронные сети.

c) Каркасные нейронные сети.

d) Криптографически защищённые нейронные сети.

e) Звёздчатые нейронные сети.

Вопрос №41

Что такое автокодировщик (Autoencoder)?

a) Нейронная сеть, способная восстанавливать исходные данные из сокращённого представления.

b) Программа для архивации файлов.

c) Микрокомпьютер для записи видео.

d) Электронный ключ доступа.

e) Система для синхронизации устройств.

Вопрос №42

Что такое архитектура GRU (Gated Recurrent Unit)?

a) Усовершенствованный вариант RNN, облегчённый аналог LSTM.

b) Метод проектирования зданий и сооружений.

c) Система управления воздушным движением.

d) Метод создания оптических приборов.

e) Организация международной конференции.

Вопрос №43

Что такое task Sentiment Analysis?

a) Задача определения эмоциональной окраски текста (позитивная, негативная, нейтральная).

b) Определить жанр литературного произведения.

c) Перевод текста на иностранный язык.

d) Извлечь названия компаний из текста.

e) Идентификация авторов текстов.

Вопрос №44

Что такое architecture AlexNet?

a) Первая победившая в конкурсе ImageNet свёрточная нейронная сеть, инициировавшая эпоху глубокого обучения.

b) Система охраны территории.

c) Проект многоэтажного здания.

d) План развития городских территорий.

e) Архитектура специального робота-помощника.

Вопрос №45

Какая архитектура сети эффективно справляется с длинными зависимостями в тексте?

a) LSTM (Long Short-Term Memory).

b) Матричная архитектура.

c) Круговая архитектура.

d) Пространственная архитектура.

e) Темпоральная архитектура.

Вопрос №46

Какая нейронная сеть используется для классификации изображений?

a) CNN (Convolutional Neural Network).

b) RNN (Recurrent Neural Network).

c) DNA (Deoxyribonucleic Acid).

d) RTF (Rich Text Format).

e) PNG (Portable Network Graphics).

Вопрос №47

Какая функция активации обычно применяется в RNN?

a) tanh (гиперболический тангенс).

b) Функция степенного роста.

c) Косинусоидальная функция.

d) Линейная функция.

e) Логарифмическая функция.

Вопрос №48

Что означает аббревиатура GRU?

a) Gated Recurrent Unit (модифицированная рекуррентная единица с управляемыми воротами).

b) Generalized Radio Unit.

c) Graphical Rendering Utility.

d) Global Routing Update.

e) Genetic Research Unit.

Вопрос №49

Что такое multi-head attention в трансформерах?

a) Несколько параллельных каналов внимания, повышающих точность обработки текста.

b) Процесс одновременного чтения нескольких книг.

c) Актёры, играющие несколько ролей одновременно.

d) Режим многопоточного рендеринга графики.

e) Совместная деятельность нескольких писателей.

Вопрос №50

Что такое vanishing gradient problem?

a) Проблема утраты значительных изменений градиента при прохождении через глубокие слои сети.

b) Утрата электрических импульсов при прохождении сигнала по проводнику.

c) Забывание человеком пройденного материала спустя длительное время.

d) Повреждение электронных плат при перегревании.

e) Старение батарей мобильных телефонов.

Тема 6 «Распределённые вычисления и параллельные алгоритмы»

Вопрос №1

Что такое MapReduce?

a) Фреймворк для параллельной обработки больших объемов данных.

b) Протокол безопасной передачи данных.

c) Ядро операционной системы Linux.

d) Язык программирования общего назначения.

e) Интернет-протокол передачи гипертекстовых документов.

Вопрос №2

Сколько этапов включает в себя алгоритм MapReduce?

a) Два этапа: Map и Reduce.

b) Три этапа: Load, Process, Save.

c) Четыре этапа: Input, Sort, Combine, Output.

d) Один этап: Transform.
e) Пять этапов: Read, Filter, Compute, Aggregate, Write.

Вопрос №3
Что выполняет mapper в MapReduce?
a) Читает данные, формирует пары ключ-значение и отправляет их на дальнейшую обработку.
b) Обеспечивает безопасность данных.
c) Организует физическую инфраструктуру сервера.
d) Ведёт учет пользователей системы.
e) Отправляет запросы пользователям.

Вопрос №4
Что такое NameNode в HDFS?
a) Центральный узел, управляющий метаданными файлов и каталогов.
b) Файловый менеджер Windows.
c) Физический сервер для хранения данных.
d) Пользовательская база данных.
e) Сервис облачного хостинга.

Вопрос №5
Что такое DataNode в HDFS?
a) Узел, хранящий реальные данные.
b) Центральный процессор.
c) Веб-сервер.
d) Клиентское приложение.
e) Графический интерфейс пользователя.

Вопрос №6
Как называется принцип записи данных в HDFS?
a) Write-once read-many (WORM).
b) Read-only access.
c) Append-only mode.
d) Multiple write policy.
e) Strict consistency protocol.

Вопрос №7
Что обеспечивает отказоустойчивость в HDFS?
a) Репликация данных.
b) Цифровая подпись.
c) Шифрование трафика.
d) Аппаратное обеспечение.
e) Авторизация пользователей.

Вопрос №8
Какая операционная система рекомендована для развертывания HDFS?
a) Unix-подобные ОС (Linux, BSD).
b) Microsoft Windows.
c) macOS.
d) Android.

e) Chrome OS.

Вопрос №9

Что происходит в фазе shuffle в MapReduce?

a) Сортировка и передача данных к соответствующим редукторам.

b) Сжатие данных для экономии места.

c) Запись данных на жесткий диск.

d) Проверка прав доступа к данным.

e) Шифрование передаваемых данных.

Вопрос №10

Какой тип задачи отлично подходит для MapReduce?

a) Параллельные массовые вычисления на больших объемах данных.

b) Интерактивные приложения реального времени.

c) Интеграция веб-сервисов.

d) Написание драйверов устройств.

e) Разработку настольных приложений.

Вопрос №11

Что такое Apache Spark?

a) Платформа для быстрой обработки больших данных с поддержкой множества задач, включая SQL-запросы, машинное обучение и потоковую обработку.

b) Система управления базами данных.

c) Графический интерфейс для аналитики данных.

d) Сервис облачного хранения данных.

e) Система автоматизации бизнес-процессов.

Вопрос №12

Что такое Driver в архитектуре Apache Spark?

a) Главный компонент, который координирует выполнение задач на рабочем узле.

b) Рабочий узел, выполняющий задачи.

c) Менеджер кластера.

d) Общий интерфейс взаимодействия с внешними сервисами.

e) Процесс планирования задач на уровне железа.

Вопрос №13

Что такое Executor в Apache Spark?

a) Рабочее приложение, исполняемое на узле Worker Nodes, выполняющее задачи.

b) Мастер-набор инструментов для администрирования кластера.

c) Центральное звено координации заданий.

d) Среда разработки для написания задач.

e) Вспомогательный инструмент для тестирования программного кода.

Вопрос №14

Что такое RDD (Resilient Distributed Dataset) в Apache Spark?

a) Иммутабельная коллекция данных, распределённая по узлам кластера.

b) Название таблицы базы данных.

c) Служба репликации данных.

d) Шаблон дизайна приложения.

e) Внешний источник данных.

Вопрос №15

Что означает термин lazy evaluation в Apache Spark?

a) Отложенное выполнение операций, происходящее только тогда, когда потребуется реальный результат.

b) Задержка выполнения из-за низкой производительности системы.

c) Неэффективная реализация вычислений.

d) Раннее завершение выполнения из-за недостатка ресурсов.

e) Ошибка выполнения программы.

Вопрос №16

Какая операция в Apache Spark относится к Transformation?

a) map

b) collect

c) count

d) takeSample

e) first

Вопрос №17

Какая операция в Apache Spark относится к Action?

a) filter

b) reduceByKey

c) join

d) collect

e) union

Вопрос №18

Что такое Spark SQL?

a) Модуль Apache Spark, поддерживающий SQL-подобные запросы для анализа структурированных данных.

b) Базовый SQL-движок MySQL.

c) Тип индексов в NoSQL-хранилищах.

d) Сторонний движок обработки Big Data.

e) SQL-консоль командной строки.

Вопрос №19

Что такое Spark Streaming?

a) Компонент Apache Spark для обработки потоков данных в реальном времени.

b) Онлайн-игровой стриминг.

c) Платформа потокового вещания видеоконтента.

d) Онлайн-конструктор презентаций.

e) Сервис доставки сообщений.

Вопрос №20

Что такое MLlib в Apache Spark?

a) Пакет машинного обучения, встроенный в Apache Spark.

b) Отдел маркетинга крупной технологической компании.

c) Система документооборота предприятия.

d) Open-source библиотека для разработки игр.

e) Популярный мессенджер.

Вопрос №21

Что такое Lineage Tracking в Apache Spark?

a) Механизм отслеживания истории происхождения данных, позволяющий восстановить данные в случае повреждения.

b) Мониторинг производительности узлов кластера.

c) Инструментарий для профилирования CPU.

d) Диагностика коммуникационных каналов.

e) Тестирование совместимости версий ПО.

Вопрос №22

Что такое Partitioning в Apache Spark?

a) Разделение данных на сегменты для распараллеливания вычислений.

b) Закрытие открытых сессий.

c) Слияние данных из разных источников.

d) Кэширование данных в памяти.

e) Запись данных на постоянное хранилище.

Вопрос №23

Что такое micro-batching в Spark Streaming?

a) Метод обработки данных небольшими партиями для приближения к реальному времени.

b) Массовая рассылка писем.

c) Установка микропакетов программ.

d) Мини-разделы веб-сайта.

e) Загрузка пакетов в облако.

Вопрос №24

Что такое resilient distributed dataset (RDD) в Apache Spark?

a) Распределённое представление данных, разбитое на части и размещённое на разных узлах кластера.

b) Виртуальная машина для запуска задач.

c) Комплексный анализ социальных медиа.

d) Серверное оборудование для установки инфраструктуры.

e) Команда разработчиков.

Вопрос №25

Какая команда в Apache Spark используется для сбора всех элементов RDD на мастер-ноде?

a) collect

b) filter

c) reduceByKey

d) groupByKey

e) sortByKey

Вопрос №26

Что такое Apache Kafka?

a) Распределённая потоковая платформа для передачи сообщений в реальном времени.

b) Универсальный протокол шифрования данных.

c) Система управления проектами Agile-методологией.

d) Консольный терминал UNIX-систем.

e) Социальная сеть профессиональных контактов.

Вопрос №27
Кто разработал Apache Kafka?
a) Компания LinkedIn.
b) Университет Стэнфорда.
c) Фонд Apache Software Foundation.
d) Институт Хассо Платтнера.
e) Компания Facebook.

Вопрос №28
Что такое Producer в Kafka?
a) Источник сообщений, отправляющий данные в топики.
b) Потребитель сообщений.
c) Часть физической инфраструктуры сервера.
d) Серверная среда для хранения файлов.
e) Тип контейнера Docker.

Вопрос №29
Что такое Consumer в Kafka?
a) Подписчик, получающий сообщения из топиков.
b) Поставщик услуг.
c) Автономный узел кластера.
d) Маршрутизатор данных.
e) Протокол сетевого взаимодействия.

Вопрос №30
Что такое Broker в Kafka?
a) Сервер Kafka, принимающий и сохраняющий сообщения.
b) Транзитный узел межсетевого взаимодействия.
c) Маркер окончания данных.
d) Внутренний реестр системы.
e) Статистика производительности.

Вопрос №31
Что такое Topic в Kafka?
a) Категория или очередь сообщений.
b) Географическое местоположение.
c) Кодировка символов.
d) Логический контейнер DNS.
e) Род занятий сотрудников.

Вопрос №32
Что такое Partition в Kafka?
a) Секция внутри топика, позволяющая увеличить пропускную способность.
b) Единичный файл данных.
c) Домашний каталог пользователя.
d) Межпроцессорная связь.
e) Регистрация пользователя.

Вопрос №33
Что такое Offset в Kafka?
a) Уникальный идентификатор сообщения в секции.
b) Порт TCP/IP.
c) Уровень приоритета задачи.
d) Размер пакета данных.
e) Дата истечения срока лицензии.

Вопрос №34
Что такое Apache Flink?
a) Фреймворк для обработки потоковых данных в реальном времени.
b) Конфигурационный файл Linux.
c) Клиент почтовой службы.
d) Языковой стандарт ISO.
e) Информационно-справочная служба поддержки клиентов.

Вопрос №35
Что такое JobManager в Apache Flink?
a) Компонент, координирующий выполнение задач.
b) Менеджер проектов в производственной среде.
c) Папка для хранения конфигурационных файлов.
d) Пользовательский профиль.
e) VPN-шлюз.

Вопрос №36
Что такое TaskManager в Apache Flink?
a) Компонент, исполняющий отдельные задачи.
b) Редактор конфигурации.
c) Балансировщик нагрузки.
d) Планировщик расписания задач.
e) Транспортный шлюз.

Вопрос №37
Что такое Source в Apache Flink?
a) Источник данных, например, Kafka или HDFS.
b) Тип маршрута маршрутизатора.
c) Файловая система NTFS.
d) Телекоммуникационный провайдер.
e) Интернет-магазин.

Вопрос №38
Что такое Sink в Apache Flink?
a) Приемник данных, куда отправляются результаты обработки.
b) Склад материалов.
c) Сеть хранения данных SAN.
d) Назначение команды на выполнение.
e) Конвейер производства.

Вопрос №39
Что такое Windowing в Apache Flink?
a) Временное окно для агрегации событий.

b) Панель навигации браузера.
c) Горизонтальная полоса прокрутки.
d) Модуль шифрования данных.
e) Правило делегирования полномочий.

Вопрос №40

Что такое Event-time processing в Apache Flink?

a) Обработка событий на основе отметки времени самого события, а не момента поступления данных.
b) Календарь корпоративных мероприятий.
c) Мониторинг цен акций.
d) Проведение опросов общественного мнения.
e) Правила внутреннего трудового распорядка.

Тема 7 «Интеграция и преобразование данных»

Вопрос №1

Что является основной целью этапа очистки и подготовки данных?

a) Повышение точности результатов анализа путем улучшения качества исходных данных.
b) Увеличение объема данных для лучшего представления модели.
c) Автоматизированное заполнение пропусков в данных.
d) Преобразование всех числовых данных в бинарные.
e) Улучшение визуализации данных.

Вопрос №2

Какой метод используется для замены пропущенных значений наиболее распространенными значениями категории?

a) Label Encoding
b) One-Hot Encoding
c) Mean Imputation
d) Mode Imputation
e) Drop NaN

Вопрос №3

Как называется процесс преобразования категорий в двоичные признаки?

a) Normalization
b) Standardization
c) Feature Scaling
d) One-Hot Encoding
e) Log Transformation

Вопрос №4

Какие методы применяются для борьбы с проблемой несбалансированности классов?

a) Overfitting и Underfitting
b) Undersampling и Oversampling
c) Min-Max Scaling и Z-Score Normalization
d) Clustering и Classification
e) Bagging и Boosting

Вопрос №5

Какой способ обработки данных применяется для изменения масштаба числовых признаков к заданному интервалу?

a) One-Hot Encoding
b) Min-Max Scaling
c) Box-Cox Transformation
d) PCA (Principal Component Analysis)
e) Regularization

Вопрос №6

Чем отличается Label Encoding от One-Hot Encoding?

a) Label Encoding присваивает каждому классу уникальное число, тогда как One-Hot создает отдельный признак для каждого класса.

b) Label Encoding сохраняет порядок классов, а One-Hot преобразует классы в случайные числа.

c) Label Encoding применим только к количественным данным, а One-Hot — к качественным.

d) Label Encoding удаляет редкие классы, а One-Hot сохраняет их.

e) Label Encoding увеличивает размер данных, а One-Hot уменьшает.

Вопрос №7

Почему обработка выбросов важна в процессе подготовки данных?

a) Выбросы увеличивают объем памяти, необходимой для хранения данных.
b) Выбросы уменьшают точность методов машинного обучения.
c) Выбросы помогают выявить скрытые закономерности в данных.
d) Выбросы делают данные менее читаемыми визуально.
e) Выбросы ускоряют обучение алгоритмов.

Вопрос №8

Какой алгоритм позволяет создать дополнительные экземпляры объектов редкого класса?

a) SMOTE (Synthetic Minority Oversampling Technique)
b) Random Forest
c) K-means clustering
d) Gradient Boosting Machines
e) Decision Tree

Вопрос №9

Какая техника нормализации масштабирует признаки таким образом, чтобы среднее значение было равно нулю, а дисперсия равнялась единице?

a) Min-Max Scaling
b) Robust Scaler
c) Standard Scaler
d) Quantile Transformer
e) Power Transformer

Вопрос №10

Зачем выполняется объединение данных из разных источников?

a) Чтобы увеличить разнообразие признаков и улучшить общую картину данных.
b) Для уменьшения размера базы данных.
c) Чтобы устранить корреляцию между признаками.
d) Для повышения скорости обработки данных.

e) Чтобы снизить вычислительную сложность алгоритма.

Вопрос №11
Из какого шага состоит первая фаза ETL-процессов?
a) Extract (Извлечение данных)
b) Transform (Трансформация данных)
c) Load (Загрузка данных)
d) Monitoring (Мониторинг данных)
e) Backup (Резервное копирование данных)

Вопрос №12
Что представляет собой второй шаг ETL-процесса?
a) Фаза архивирования старых данных
b) Процесс объединения данных
c) Stage Load (загрузка временных данных)
d) Трансформация данных
e) Incremental Loading (инкрементальная загрузка)

Вопрос №13
Какой метод чаще всего используется для первичной обработки данных в фазе трансформации?
a) Min-Max Scaling
b) Fillna() / impute missing data
c) Change-data capture
d) Full extract
e) Real-time streaming

Вопрос №14
Что обозначает термин "Batch loading"?
a) Полная перезагрузка данных из источника
b) Частичная загрузка новых данных
c) Загрузка данных небольшими порциями
d) Прямая загрузка данных в таблицы
e) Инкрементальная выгрузка устаревших данных

Вопрос №15
Какой тип данных преимущественно обрабатывается в ETL-процессе?
a) Числовые и временные данные
b) Только текстовые данные
c) Только мультимедийные файлы
d) Географические координаты
e) Все типы данных

Вопрос №16
В каком инструменте удобно создавать пайплайны ETL?
a) Apache Spark
b) Tableau
c) Power BI
d) Apache Airflow
e) Jupyter Notebook

Вопрос №17
Какая операция относится к стадии трансформации данных?
a) Join (объединение данных)
b) Upload (загрузка данных)
c) Delete (удаление данных)
d) Query (запрос данных)
e) Restore (восстановление данных)

Вопрос №18
Что значит термин "incremental loading"?
a) Постепенная загрузка небольших объемов данных
b) Периодическая полная загрузка данных
c) Одновременная загрузка всех данных
d) Формат сериализации данных
e) Логирование действий над данными

Вопрос №19
Какой компонент входит в структуру классического пайплайна ETL?
a) Testing (тестирование)
b) Logging (ведение журнала)
c) Optimization (оптимизация)
d) Security (безопасность)
e) Transform (преобразование)

Вопрос №20
Какая задача решается на этапе Load (загрузки)?
a) Размещение данных в центральное хранилище
b) Коррекция ошибок в данных
c) Выбор оптимального формата данных
d) Поиск дублирующих записей
e) Генерация отчетов

Вопрос №21
Что такое "change-data capture"?
a) Метод полного извлечения данных
b) Способ отслеживания изменений данных
c) Техника интеграции с REST API
d) Технологию прямого чтения данных
e) Тип оптимизации запросов

Вопрос №22
Какой инструмент поддерживает эффективное управление и автоматизацию пайплайнов ETL?
a) Google Sheets
b) Microsoft Word
c) Adobe Photoshop
d) Talend Open Studio
e) Skype for Business

Вопрос №23

Какой этап ETL отвечает за преобразование данных из сырого состояния в нужный формат?

a) Extract (извлечение)
b) Transform (трансформация)
c) Load (загрузка)
d) Monitor (мониторинг)
e) Optimize (оптимизация)

Вопрос №24

В какой форме лучше представлять данные на выходе ETL-процессов?
a) Исходные необработанные данные
a) Raw format (необработанная форма)
b) Cleaned and structured form (очищенная и структурированная форма)
c) Binary format (двоичный формат)
d) Compressed files (архивированные файлы)
e) Decrypted format (расшифрованный формат)

Вопрос №25

Что означает термин "Full extract"?
a) Получение только части данных из источника
b) Полное извлечение всех необходимых данных
c) Использование только обновленных данных
d) Резервное копирование данных
e) Реорганизация существующих данных

Вопрос №26

Какой механизм используют для обнаружения и обработки изменений в данных?
a) Change-data capture
b) Incremental loading
c) Batch processing
d) Direct loading
e) Cross validation

Вопрос №27

Какой принцип важен при проектировании устойчивого пайплайна ETL?
a) Минимизация ручного труда
b) Максимальная нагрузка на систему
c) Отсутствие резервного копирования
d) Игнорирование стандартов индустрии
e) Максимально сложная архитектура

Вопрос №28

Какой компонент обеспечивает перенос данных из одного места в другое?
a) Analytics Engine
b) Storage System
c) Transport Layer
d) Metadata Repository
e) Dashboard Interface

Вопрос №29

Какой шаг предполагает проверку целостности данных и выявление ошибок?
a) Aggregation
b) Validation
c) Extraction
d) Encryption
e) Compression

Вопрос №30
Какой инструмент подходит для визуализации сложных процессов ETL?
a) Notepad++
b) GitHub Desktop
c) MS Paint
d) Visio или Lucidchart
e) Google Docs

Вопрос №31
Какая особенность характерна для формата JSON?
a) Поддерживает исключительно двоичное представление данных.
b) Требует строгого соблюдения иерархической структуры данных.
c) Легко читается человеком благодаря простоте синтаксиса.
d) Используется преимущественно для описания изображений и мультимедиа.
e) Не позволяет хранить числовые значения.

Вопрос №32
Что означает термин "полуструктурированные данные"?
a) Данные представлены исключительно в форме таблиц.
b) Данные содержат фиксированную схему и обязательные атрибуты.
c) Данные обладают некоторой структурой, но допускают вариативность в представлении.
d) Это синоним термина "неструктурированные данные".
e) Данные представляют собой чистый текст без какой-либо разметки.

Вопрос №33
Какие недостатки имеет формат CSV?
a) Высокая скорость чтения и записи данных.
b) Возможность представлять сложные типы данных, включая вложенные объекты.
c) Отсутствие поддержки для выраженной вложенности и сложных типов данных.
d) Автоматическое распознавание типа каждого столбца.
e) Оптимальная поддержка Unicode символов.

Вопрос №34
Чем отличается формат AVRO от JSON и XML?
a) AVRO не предназначен для хранения данных.
b) AVRO поддерживает только строковые типы данных.
c) AVRO обеспечивает эффективное бинарное представление данных и возможность сжатия.
d) AVRO работает исключительно с полуструктурированными данными.
e) AVRO применяется исключительно в мобильных приложениях.

Вопрос №35

Какой инструмент предпочтителен для быстрого преобразования XML-документов в JSON?

a) OpenCV
b) PyYAML
c) pandas
d) xmltodict
e) jq

Вопрос №36

В каком формате удобно передавать большие объемы структурированных данных с минимальными затратами памяти?

a) JSON
b) XML
c) CSV
d) AVRO
e) PNG

Вопрос №37

Почему формат CSV менее универсален по сравнению с JSON и XML?

a) CSV файлы быстрее обрабатываются современными компьютерами.
b) CSV формат лучше всего подходит для графического отображения данных.
c) CSV не поддерживает вложенные структуры и типы данных сложнее простых значений.
d) CSV автоматически проверяет тип данных в каждой ячейке.
e) CSV файл легче сжимается архиваторами.

Вопрос №38

Какие преимущества имеет использование библиотек при работе с большими наборами данных?

a) Увеличиваются временные затраты на обработку данных.
b) Повышается риск ошибок из-за ручной реализации алгоритмов.
c) Упрощается работа с данными путем автоматизации распространенных операций.
d) Библиотеки ограничены возможностями конкретных платформ и операционных систем.
e) Использование библиотек снижает производительность приложений.

Вопрос №39

Когда целесообразно применять частичную обработку данных перед полным преобразованием?

a) Если требуется сохранить весь исходный набор данных без изменений.
b) При наличии большого объема данных, среди которых значительная часть избыточна.
c) Только тогда, когда используются устаревшие технологии и инструменты.
d) Частичная обработка увеличивает вероятность потери важных данных.
e) Всегда, независимо от размера набора данных.

Вопрос №40

Что такое "проблема несоответствия схем данных" ("schema mismatch")?

a) Ошибка, возникающая при неправильном кодировании текста.

b) Несоответствие структуры данных между источником и приемником данных.
c) Проблема недостаточной производительности компьютера.
d) Процесс автоматического выбора наилучшего способа сохранения данных.
e) Невозможность сохранять большие объемы данных одновременно.

Вопрос №41
Какой этап необходим при конвертации JSON в CSV?
a) Перевод всех чисел в строки.
b) Расширение файла ".json" на ".csv".
c) Преобразование вложенных объектов в плоскую таблицу.
d) Создание резервной копии данных в облаке.
e) Удаление дубликатов записей вручную.

Вопрос №42
Что подразумевает "частично организованные данные"?
a) Полностью структурированный набор данных, соответствующий определенной схеме.
b) Данные, содержащие смесь организованной и случайной структуры.
c) Формат данных, предназначенный исключительно для изображений.
d) Совершенно неорганизованный набор данных.
e) Термин, обозначающий отсутствие данных вообще.

Вопрос №43
Что представляет собой проблема отсутствия данных при трансформации форматов?
a) Наличие одинаковых данных в разных источниках.
b) Некорректное применение шаблона при обработке файлов.
c) Информация отсутствует в источнике, вызывая ошибку при сопоставлении полей.
d) Возникает конфликт между несколькими версиями программного обеспечения.
e) Пользователь случайно удалил важные данные.

Вопрос №44
Какая стратегия помогает ускорить процесс преобразования данных?
a) Избегайте использования существующих библиотек и пишите собственные алгоритмы.
b) Используйте примитивные средства редактирования текста, такие как Блокнот.
c) Применяйте кэширование результатов повторяющихся операций.
d) Регулярно перезагружайте компьютер для освобождения ресурсов.
e) Работайте только в однопоточном режиме для упрощения разработки.

Вопрос №45
Что из перечисленного относится к методикам оптимизации при преобразовании данных?
a) Добавление новых функций в программное обеспечение без тестирования.
b) Разработка собственного уникального формата данных.
c) Кэширование обработанных данных для последующего повторного использования.
d) Постоянное изменение схемы данных без уведомления пользователей.

e) Исключение возможности масштабируемости решений.

Тема 8 «Анализ поведения пользователей и персонализация услуг»

Вопрос № 1

Что является главной целью профилирования клиентов?

a) Повышение эффективности производства.

b) Разработка индивидуальной стратегии взаимодействия с клиентами.

c) Увеличение скорости доставки товаров.

d) Оптимизация логистических процессов.

e) Уменьшение затрат на рекламу.

Вопрос № 2

Какие виды данных используются для профилирования клиентов?

a) Только демографические данные.

b) Демографические, географические, покупательские истории и психологические факторы.

c) Только финансовые отчёты организаций.

d) Информация о конкурентах.

e) Рейтинги сотрудников организации.

Вопрос № 3

Какой метод анализа позволяет разделить клиентов по давности последней покупки, частоте покупок и сумме трат?

a) ABC-анализ.

b) K-медианы.

c) Чередование оценивания (A/B-тестирование).

d) RFM-анализ.

e) Метод главных компонент.

Вопрос № 4

Как называется метод, позволяющий выделить группу клиентов с одинаковым финансовым вкладом в доходы компании?

a) K-кластеры.

b) Ассоциативный анализ.

c) Регрессия.

d) ABC-анализ.

e) Линейное программирование.

Вопрос № 5

Для какого метода анализа характерно случайное назначение начальных центров кластеров?

a) Байесовская классификация.

b) Нейронные сети.

c) K-means.

d) Аналитическое ранжирование.

e) Дискриминантный анализ.

Вопрос № 6

Что такое временная серия в поведенческом анализе?

a) Последовательность событий в пространстве.

b) График производительности предприятия.

c) Географическое распределение клиентов.
d) Набор данных, упорядоченный по времени.
e) Прогноз роста рынка.

Вопрос № 7

Чем характеризуется когорта в поведенческом анализе?
a) Группа клиентов, совершающих одинаковые покупки одновременно.
b) Совокупность продуктов одной категории.
c) Группы пользователей с разным уровнем доходов.
d) Клиенты, зарегистрировавшиеся в одну эпоху времени.
e) Компании-конкуренты в отрасли.

Вопрос № 8

Что изучается в анализе воронки продаж?
a) Пути клиентов от знакомства с продуктом до покупки.
b) Особенности производственных процессов.
c) Эффективность кадровой политики.
d) Изменчивость цен на товары.
e) Способы привлечения новых поставщиков.

Вопрос № 9

Какой показатель отражает процент клиентов, остающихся активными спустя определённое время?
a) Конверсионный коэффициент.
b) Retention Rate.
c) Индекс удовлетворенности.
d) Net Promoter Score.
e) Customer Acquisition Cost.

Вопрос № 10

Какой алгоритм классификации чаще всего применяется для разделения клиентов на группы?
a) Линейная регрессия.
b) Случайный лес.
c) Корреляционный анализ.
d) Обобщённый линейный дискриминант.
e) Техническое обслуживание (maintenance).

Вопрос № 11

Какой алгоритм кластеризации используется для выделения однородных групп клиентов?
a) Логистическая регрессия.
b) Деревья решений.
c) Квадратичное программирование.
d) K-means.
e) Генерализация методов Монте-Карло.

Вопрос № 12

Что позволяет спрогнозировать поведение клиентов на основе исторических данных?
a) Методы имитации хаоса.
b) Экспертные оценки.

c) Классические экономические теории.
d) Машинное обучение.
e) Традиционная статистика.

Вопрос № 13

Какая технология используется для обработки больших объемов данных в аналитических системах?

a) MS Excel.
b) IBM SPSS.
c) PowerPoint.
d) Hadoop/HDFS.
e) AutoCAD.

Вопрос № 14

Какой инструмент Python широко используется для работы с массивами данных?

a) PyGame.
b) Django.
c) Matplotlib.
d) Pandas.
e) BeautifulSoup.

Вопрос № 15

Какой инструмент позволяет строить графики и диаграммы для визуализации данных?

a) Google Docs.
b) Microsoft Word.
c) Apache Spark.
d) Jupyter Notebook.
e) Tableau Public.

Вопрос №16

Какой тип коллаборативной фильтрации основан на сравнении пользователей?

a) Item-Based CF.
b) User-Based CF.
c) Context-Aware CF.
d) Hybrid CF.

Вопрос №17

Какая метрика часто используется для расчёта расстояния между пользователями в коллаборативной фильтрации?

a) Евклидово расстояние.
b) Махаланобисово расстояние.
c) Коэффициент Пирсона.
d) Минковский метр.
e) Манхэттенская дистанция.

Вопрос №18

Что является основным недостатком метода коллаборативной фильтрации?

a) Простота реализации.
b) Высокая точность предсказания.

c) Cold Start Problem (проблема холодного старта).
d) Низкая вычислительная сложность.
e) Широкий спектр применимости.

Вопрос №19

Чем отличается метод рекомендаций на основе контента от коллаборативной фильтрации?

a) Использует оценки других пользователей.
b) Требует много ресурсов для хранения данных.
c) Оперирует признаками самого объекта.
d) Применяется исключительно в онлайн-магазинах.
e) Всегда даёт точные рекомендации.

Вопрос №20

Почему рекомендации на основе контента могут быть менее разнообразными?
a) Они учитывают мнения всех пользователей.
b) Пользователь видит слишком много предложений.
c) Зависят лишь от предыдущих выборов пользователя.
d) Метод работает медленно.
e) Этот подход редко применяется в индустрии.

Вопрос №21

Какие преимущества имеет подход Content-Based перед Collaborative Filtering?
a) Не требует сбора большого количества оценок
b) Предсказуемые и понятные рекомендации.
c) Отсутствие Cold Start Problem.
d) Лучшая масштабируемость на большие объёмы данных.
e) Более точное выявление редких предпочтений.

Вопрос №22

Что такое гибридный подход в рекомендациях?
a) Использование только одной модели для выдачи рекомендаций.
b) Полностью игнорирует характеристики объектов.
c) Совмещение методов Content-Based и Collaborative Filtering.
d) Работает только с популярными категориями продуктов.
e) Учитывает исключительно поведение пользователя.

Вопрос №23

Как называется проблема, возникающая при отсутствии достаточного количества оценок новых пользователей или предметов?

a) Hot Start Problem.
b) Overfitting.
c) Cold Start Problem.
d) Underfitting.
e) Data Leakage.

Вопрос №24

Какой подход лучше подходит для предоставления неожиданных рекомендаций?
a) Content-Based Recommendation.
b) Collaborative Filtering.
c) Demographic Filtering.

d) Knowledge-Based Recommender System.
e) Popularity-Based Recommender System.

Вопрос №25
Для чего применяются рекомендательные системы?
a) Только для продвижения рекламы.
b) Исключительно для улучшения поисковых алгоритмов.
c) Повышение вовлечённости пользователей и улучшение качества сервиса.
d) Обеспечение безопасности транзакций.
e) Оптимизация загрузки серверов.

Вопрос №26
Что означает понятие "персонализированный маркетинг"?
a) Массовая рассылка одинаковых сообщений клиентам.
b) Реклама, ориентированная на узкую целевую аудиторию.
c) Один общий шаблон обращения для всех потребителей.
d) Создание уникального предложения для каждого отдельного клиента.
e) Акцент на массовые скидки и распродажи.

Вопрос №27
Какой метод сегментации предполагает разделение аудитории по таким характеристикам, как уровень дохода, род занятий и образовательный уровень?
a) Психографический сегмент.
b) Географический сегмент.
c) Демографический сегмент.
d) Поведенческий сегмент.
e) Экономический сегмент.

Вопрос №28
Какая формула описывает алгоритм расчета степени сходства между двумя пользователями в коллаборативной фильтрации?
a) Средневзвешенное значение рейтингов.
b) Евклидова норма.
c) Коши-Буняковского неравенство.
d) Корреляционный коэффициент Пирсона.
e) Метрика Манхэттена.

Вопрос №29
Что представляет собой поведенческая сегментация аудитории?
a) Деление аудитории по демографическим данным.
b) Разделение аудитории по уровню доходов.
c) Определение сегментов по интересам и хобби.
d) Распределение пользователей по географическому принципу.
e) Группа аудиторий, сформированных исходя из поведения потребителей (например, частота покупок, реакция на акционные предложения).

Вопрос №30
Какой критерий важен при подборе целевой аудитории для таргетированной рекламы?
a) Любовь к определённым музыкальным исполнителям.
b) Политические взгляды пользователей.

c) Социально-экономическое положение аудитории.
d) Национальность пользователей.
e) Частота посещения кинотеатра

Вопрос №31
Что показывает показатель Conversion Rate в рекламных кампаниях?
a) Число показов объявлений.
b) Средняя продолжительность просмотра рекламы.
c) Процент перехода на сайт после просмотра рекламы.
d) Общий объем трафика на сайте.
e) Доля привлечённых клиентов среди всех взаимодействовавших с объявлением.

Вопрос №32
Какие методы анализа данных помогают выявлять связи между поведением пользователей?
a) Деревья решений и логистическая регрессия.
b) Ассоциации и кластеры данных.
c) Байесовская статистика и линейная регрессия.
d) Алгоритмы машинного обучения и ассоциативные правила.
e) Многомерный статистический анализ и хронометраж действий.

Вопрос №33
Что обозначает аббревиатура ROI в маркетинге?
a) Размер рынка конкретной отрасли.
b) Эффективность работы отдела продаж.
c) Рост доли рынка компании.
d) Индекс потребительской активности.
e) Возвращённая прибыль на вложенные инвестиции.

Вопрос №34
Какая платформа предназначена специально для размещения таргетированной рекламы в русскоязычной социальной сети?
a) Google AdWords.
b) MyTarget.
c) Yandex.Direct.
d) Facebook Ads Manager.
e) VK Target.

Тема 9 «Прогнозирование и принятие решений»
Вопрос №1
Что такое регрессия в статистике?
a) Процесс формирования гипотез о будущем событии.
b) Способ разделения данных на классы.
c) Подбор функций, отражающих закономерности между переменными.
d) Вычисление средних значений в выборке.
e) Графическое отображение распределения величин.

Вопрос №2
Какая задача решается простой линейной регрессией?
a) Нахождение среднего арифметического ряда чисел.

b) Поиск прямой линии, наилучшим образом аппроксимирующей точки на плоскости.

c) Расчёт стандартного отклонения набора данных.

d) Классификация данных по категориям.

e) Выявление аномалий в датасете.

Вопрос №3

Что означает выражение “метод наименьших квадратов” (OLS)?

a) Выбор минимальной суммы квадратов отклонений точек от выбранной линии.

b) Максимизация точности модели путем добавления большего числа признаков.

c) Минимизация разницы между максимальным и минимальным значением.

d) Исключение выбросов из исходного набора данных.

e) Упрощение сложной модели до линейной формы.

Вопрос №4

Какова основная идея полиномиальной регрессии?

a) Введение степеней независимых переменных для описания нелинейных зависимостей.

b) Увеличение числа классов для лучшей классификации данных.

c) Применение дерева решений для повышения точности модели.

d) Переход от категориальных данных к численным.

e) Преобразование признаков в бинарные величины.

Вопрос №5

Когда применяется логистическая регрессия?

a) Если зависимая переменная непрерывна и нормально распределена.

b) Если зависимая переменная принимает всего два возможных значения.

c) Если нужно построить нелинейную кривую зависимости.

d) Если имеются пропущенные значения в данных.

e) Если необходимо сгладить выбросы в наборе данных.

Вопрос №6

Что такое древовидная модель в контексте регрессионного анализа?

a) Модель, использующая иерархию признаков для принятия решения.

b) Матрица корреляций между всеми парами признаков.

c) Метод визуализации плотности распределения данных.

d) Нейронная сеть с глубоким уровнем слоев.

e) Классический способ нормализации данных.

Вопрос №7

Что означают гиперпараметры в контексте регрессионных моделей?

a) Параметры, заданные вручную для настройки модели.

b) Переменные, полученные в процессе тренировки модели.

c) Данные, используемые для тестирования модели.

d) Значения, определяемые автоматически моделью.

e) Дополнительные столбцы, созданные искусственно.

Вопрос №8

Зачем применяют регуляризацию Ridge и Lasso в регрессионных моделях?

a) Чтобы увеличить число признаков в модели.

b) Чтобы уменьшить риск переобучения модели.

c) Для ускорения процесса обучения модели.
d) Для изменения шкалы признаков.
e) Для преобразования категориальных признаков в численные.

Вопрос №9

Какой показатель измеряет среднюю квадратичную ошибку (MSE) в регрессионных моделях?

a) Отклонение прогнозируемых значений от фактических.
b) Степень разброса наблюдаемой выборки вокруг среднего значения.
c) Скорость уменьшения ошибок с увеличением размера обучающей выборки.
d) Вероятность правильной классификации данных.
e) Средний размер пробелов в рядах данных.

Вопрос №10

Как называется подход, при котором выводы основаны на опыте прошлых наблюдений и реальных данных?

a) Эмпирический подход.
b) Теоретико-экспериментальный подход.
c) Байесовский подход.
d) Абстрактно-математический подход.
e) Дедуктивный подход.

Вопрос №11

Что такое временный ряд?

a) Последовательность измерений, сделанная случайно в разное время.
b) Набор чисел, расположенных хаотично.
c) Ряд показателей, зафиксированных в хронологическом порядке через равные интервалы времени.
d) Таблица значений переменных, собранных вне привязки ко времени.
e) Последовательность любых событий, произошедших подряд.

Вопрос №12

Как называют свойство временного ряда, при котором его среднее значение и дисперсия остаются постоянными во времени?

a) Цикличность.
b) Тренд.
c) Стационарность.
d) Сезонность.
e) Белошумность.

Вопрос №13

Что означает тренд в контексте анализа временных рядов?

a) Периодичность изменений, связанная с календарём.
b) Непредсказуемые скачки значений.
c) Долгосрочное устойчивое изменение уровня ряда.
d) Резкое падение или рост данных.
e) Наличие структурных разрывов.

Вопрос №14

Какая операция применяется для удаления тренда из временного ряда?

a) Дифференцирование.
b) Скользящее среднее.
c) Эксцесс.
d) Индексация.
e) Гомоскедастичность.

Вопрос №15
Что представляют собой остатки в разложении временного ряда?
a) Повторяющиеся паттерны.
b) Случайные компоненты, оставшиеся после выделения тренда и сезонности.
c) Часть данных, подверженная систематической ошибке.
d) Постоянный тренд в ряде.
e) Прогрессивно увеличивающаяся амплитуда.

Вопрос №16
Что характеризует автокорреляцию в контексте временных рядов?
a) Связь между разными переменными.
b) Волатильность рынка.
c) Шум в сигнале.
d) Корреляцию текущего значения ряда с предыдущими значениями.
e) Зависимость от внешней среды.

Вопрос №17
Что обозначает буква 'I' в сокращении ARIMA?
a) Интегральная.
b) Инновационная.
c) Интерполирующая.
d) Итерация.
e) Интерактивная.

Вопрос №18
Какая модель учитывает сезонность в прогнозировании временных рядов?
a) MA (скользящее среднее).
b) AR (авторегрессия).
c) ARIMA.
d) SARIMA.
e) VAR (векторная авторегрессия).

Вопрос №19
Какая библиотека была разработана Facebook для прогнозирования временных рядов?
a) PyTorch.
b) TensorFlow.
c) Keras.
d) Prophet.
e) Scikit-Learn.

Вопрос №20
Какая нейронная сеть эффективна для учёта долговременных зависимостей в данных временных рядов?

a) Fully Connected Neural Network (FCNN).
b) Convolutional Neural Networks (CNNs).
c) Long Short-Term Memory (LSTM).
d) Autoencoder.
e) Generative Adversarial Networks (GANs).

Вопрос №21

Что такое принятие управленческих решений?
a) Действие руководителя, направленное на повышение прибыли предприятия.
b) Автоматический процесс управления предприятием.
c) Результат компьютерного моделирования бизнеса.
d) Активность персонала по самостоятельному принятию важных решений.
e) Процесс выбора оптимального курса действий руководителем на основе доступной информации.

Вопрос №22

Что позволяет сделать анализ данных в контексте управленческих решений?
a) Определять цели и задачи организации без участия руководителей.
b) Идентифицировать скрытые закономерности и прогнозировать развитие событий.
c) Игнорировать внешние факторы и сосредоточиться на внутренних проблемах.
d) Снижать необходимость в квалифицированном персонале.
e) Привлекать инвесторов без учета рисков.

Вопрос №23

Из каких этапов состоит стандарт CRISP-DM?
a) Сбор данных → Планирование → Тестирование → Завершение проекта.
b) Постановка задачи → Анализ данных → Подготовка данных → Моделирование → Оценка → Внедрение.
c) Исследование рынка → Запуск продукта → Продажа → Поддержка клиентов.
d) Концептуализация → Проектирование → Производство → Маркетинг.
e) Определение стратегии → Организация команды → Выполнение работ → Оценка результата.

Вопрос №24

Что относится к диагностической аналитике?
a) Установление текущих показателей деятельности компании.
b) Генерация прогнозов развития событий на основании исторических данных.
c) Формулирование советов по улучшению существующих бизнес-процессов.
d) Понимание причин возникновения проблем или успехов.
e) Автоматическое формирование отчетов о проделанной работе.

Вопрос №25

Что представляет собой SWOT-анализ?
a) Способ проверки соответствия продукции стандартам качества.
b) Анализ финансового состояния компании.
c) Определение оптимальной численности персонала.
d) Рассмотрение стратегических позиций компании через сильные и слабые стороны, возможности и угрозы.
e) Метод исследования конкурентоспособности технологий.

Вопрос №26
Что понимается под группой "A" в ABC-анализе?
a) Товары или клиенты с низкой прибылью.
b) Стратегически важные продукты или клиенты с наибольшей долей прибыли.
c) Объекты с низкими показателями риска.
d) Перспективные рынки сбыта.
e) Категория изделий с самыми большими затратами на производство.

Вопрос №27
Как классифицируются товары в матрице БКГ ("звёзды")?
a) Высокие темпы роста рынка и небольшие рыночные доли.
b) Медленный темп роста и высокий контроль рынка.
c) Небольшие перспективы роста и низкие доли рынка.
d) Незначительные продажи и минимальные затраты на поддержку.
e) Быстрый рост рынка и лидирующее положение на нём.

Вопрос №28
Что позволяет сократить PCA (анализ главных компонент)?
a) Количество используемых моделей для анализа данных.
b) Объем оперативной памяти компьютера.
c) Время выполнения операций сотрудниками компании.
d) Необходимость ручного ввода данных.
e) Размеры многомерных массивов данных, сохраняя большую часть вариации.

Вопрос №29
Какой метод используется для оптимизации складских запасов?
a) Управление персоналом методом кадрового резерва.
b) Сокращение ассортимента продукции.
c) Генетические алгоритмы.
d) Принцип Парето (правило 80/20).
e) Оптимизация уровней запаса на основе статистики заказов и расходов.

Вопрос №30
Что такое Tableau в контексте анализа данных?
a) Язык программирования для научных исследований.
b) Универсальная база данных для компаний малого бизнеса.
c) Сервис для автоматического перевода финансовой отчетности.
d) Интернет-платформа для онлайн-коммуникаций внутри коллектива.
e) Программа для интерактивной визуализации и анализа данных.

Задачи для решения на Python

Тема 1 «Основные понятия и концепции анализа больших данных»

Задача №1. Обработка и анализ огромного массива данных о пользователях соцсетей

Описание задачи: Социальная сеть собрала большой массив данных о миллионах пользователей. Необходимо проанализировать активность пользователей по следующим показателям: средняя длина постов, среднее количество лайков и комментариев на посты, а также общее количество публикаций за неделю. Вам предоставлены файлы формата CSV объемом около 1 ГБ с миллионами записей. Напишите скрипт на Python, используя библиотеки pandas и numpy, который

произведёт необходимые расчеты и выведет отчёты.

Задача №2. Анализ Big Data в медицине: классификация пациентов по симптомам заболевания

Описание задачи: Имеется огромный массив медицинских данных, содержащий записи о пациентах и симптомы заболеваний. Необходимо разработать классификационную модель на основе алгоритмов машинного обучения, позволяющую точно различать пациентов с заболеваниями сердца и диабетом. Используйте фреймворк scikit-learn.

Задача №3. Работа с потоковыми данными реального времени с использованием Apache Kafka

Описание задачи: Создать программу на Python, принимающую данные из потока (Apache Kafka) и производящую агрегацию этих данных в режиме реального времени. Данные поступают непрерывно и содержат следующую информацию: время публикации, уникальный ID пользователя, текст публикации, количество лайков и комментариев. Агрегируйте количество уникальных авторов публикаций за последний час.

Задача №4. Распознавание изображений на основе свёрточных нейронных сетей (CNN) для классификации фотографий зданий и природных ландшафтов

Описание задачи: Необходимо создать CNN-модели для распознавания типов изображений: фотографии зданий и природные пейзажи. Используйте фреймворк TensorFlow. Входные данные состоят из файлов JPEG, размещённых в директориях buildings и landscape.

Задача №5. Применение Natural Language Processing (NLP) для анализа тональности отзывов

Описание задачи: Имеется набор текстовых данных с отзывами пользователей о продукте. Необходимо создать NLP-классификатор, который сможет анализировать отзывы и выдавать оценку: положительный отзыв или отрицательный. Использовать фреймворк transformers и предварительно подготовленную модель трансформеров.

Задача №6. Разработка рекомендательной системы на основе анализа предпочтений пользователей

Описание задачи: Разработать рекомендательную систему, которая на основе предыдущего опыта пользователей выдает рекомендации товаров, наиболее вероятно привлекающие конкретного пользователя. Используйте технику коллаборативной фильтрации (surprise).

Задача №7. Применение Hadoop MapReduce для обработки огромных объемов данных

Описание задачи: Разработайте простую реализацию MapReduce на Python для подсчета частоты появления конкретных слов в огромном файле текстовых документов. Эта реализация должна обрабатывать файл размером больше 10 ГБ.

Задача №8. Обнаружение аномалий в сетевых данных с помощью машинного обучения

Описание задачи: Определите необычные и подозрительные IP-адреса на основе журнала событий сервера, состоящего из миллиардов записей. Используется модель изолирующего леса (IsolationForest) для обнаружения аномалий.

Задача №9. Использование Spark для обработки больших графов

Описание задачи: Реализуйте программу на Python с использованием Spark GraphFrames для поиска сообщества узлов (вершин графа), принадлежащих одной группе. Дан огромный граф пользователей социальной сети с миллиардами ребер и десятков миллионов узлов.

Задача №10. Проведение A/B-тестирования на Big Data платформе

Описание задачи: Провести A/B тестирование новой версии интерфейса веб-

сайта, чтобы убедиться, что новая версия действительно увеличивает конверсию. Используйте средства агрегирования данных и статистический анализ с помощью Python и scipy.stats.

Тема 2 «Архитектура хранилищ и платформ хранения данных»

Задача №1. Работа с HDFS через Python

Описание задачи: Изучить процесс загрузки данных в HDFS и чтение их обратно с помощью Python. Используя библиотеку hdfs3, напишите код, который загружает файл в HDFS-кластер и выводит содержимое файла на экран.

Задача №2. Спарк + SQL-запросы

Описание задачи: Научитесь запускать запросы SQL поверх больших данных с использованием Apache Spark. Используя API Spark SQL, создайте таблицу на основе данных в формате Parquet и выполните запрос, возвращающий топ-10 пользователей по количеству лайков.

Задача №3. Разработка ETL-процесса в Cassandra

Описание задачи: Реализовать ETL-процесс, собирающий данные из базы данных Cassandra и формирующий агрегированную статистику по ключевым показателям. Разработайте сценарий на Python, который извлекает данные из таблицы Cassandra, рассчитывает среднее значение некоторых полей и сохраняет результат обратно в базу.

Задача №4. Анализ данных с применением MongoDB

Описание задачи: Осуществить работу с большими JSON-данными с использованием MongoDB и Python. Нужно собрать данные из коллекции MongoDB, применить фильтрацию и группировку, затем вывести статистику по определенной атрибутивной совокупности.

Задача №5. Работа с AWS S3

Описание задачи: Научиться создавать резервные копии данных и хранить их в облаке с использованием Amazon Web Services (AWS S3). Напишите Python-код, который загружает файл в корзину (bucket) на сервисе AWS S3 и проверяет доступность файла.

Задача №6. Работа с виртуальными машинами на Google Cloud Platform

Описание задачи: Организовать доступ к ресурсам виртуальной машины на Google Cloud Platform и запустить простейший Python-код. Создайте экземпляр виртуальной машины на GCP, установите интерпретатор Python и запустите сценарий на выполнение.

Задача №7. Хранение и обработка временных рядов с Cassandra

Описание задачи: Настроить хранение временных рядов данных и организовать их обработку средствами Cassandra. Необходимо реализовать систему сохранения временных рядов в базу данных Cassandra и произвести расчет средней температуры за указанный промежуток времени.

Задача №8. Организация аналитического конвейера с использованием Apache Spark Streaming

Описание задачи: Создать систему потоковой обработки данных, поступающих в реальном времени, и обработать их с помощью Spark Streaming. Используя Apache Spark, разработайте систему, которая получает данные из потока (например, Kafka) и производит базовую агрегацию, выводя результаты на консоль.

Задача №9. Машинное обучение на больших данных с использованием Google Cloud AI Platform

Описание задачи: Реализовать инфраструктуру машинного обучения на облачном сервисе Google Cloud Platform (AI Platform). Создайте проект на GCP, используйте облачную платформу AI Platform для запуска модели ML на крупномасштабных данных и разверните готовую модель для дальнейшего использования.

Задача №10. Работа с Yandex.Cloud Storage

Описание задачи: Получить опыт работы с облачными хранилищами Yandex.Cloud и освоить операции чтения и записи данных. Создайте объектный сторадж на Yandex.Cloud, подключитесь к нему и выполните базовые операции: загрузите файл, прочитайте его и получите информацию о содержимом.

Тема 3 «Статистический анализ и визуализация данных»

Задача №1. Корреляционный анализ и визуализация связей между показателями экономики

Описание задачи: Провести корреляционный анализ между ВВП, уровнем безработицы и средней заработной платой по странам мира, создать тепловые карты корреляций и проанализировать полученные выводы. Используется открытый датасет world_economy.csv, содержащий статистику стран по годам (колонки: страна, год, ВВП, уровень безработицы, средняя зарплата).

Задания:

Загрузите и исследуйте датасет.

Подготовьте данные: устраните пропуски, нормализуйте шкалу зарплат и уровней безработицы.

Вычислите матрицу корреляций между всеми тремя показателями и постройте тепловые карты.

Интерпретируйте результаты и сделайте вывод о взаимосвязях между показателями.

Задача №2. Проверка гипотез о среднем доходе населения разных регионов

Описание задачи: Проверить нулевую гипотезу о равенстве средних доходов жителей двух регионов против альтернативной гипотезы о неравенстве средних значений. Имеются две группы данных о доходах жителей региона A (region_A_income) и региона B (region_B_income).

Задания:

Загрузите и исследуйте данные.

Проверьте гипотезу о равенстве средних значений дохода методом t-теста.

Сделайте вывод о справедливости нулевой гипотезы.

Задача №3. Визуализация распределений данных с помощью box-plot'ов

Описание задачи: Исследовать асимметрию и наличие выбросов в распределениях данных, провести сравнение плотности распределения. Исходный датасет sales_data.csv, содержащий продажи продуктов в разные месяцы.

Задания:

Исследуйте исходные данные и проведите предварительную обработку.

Постройте боксплот распределения продаж по каждому продукту.

Выделите продукты с наибольшим числом выбросов и определите причины возникновения аномалий.

Задача №4. Визуализация временных рядов и прогнозирование будущего тренда

Описание задачи: Создать графики изменения цен на акции крупной компании, сделать базовый прогноз на следующий период. Используются исторические цены акций компании X в файле stock_prices.csv.

Задания:

Проведите визуализацию временного ряда цен на акции.

Примените простую скользящую среднюю для сглаживания колебаний графика.

Спрогнозируйте цену акций на следующий квартал, используя тренд.

Задача №5. Анализ факторов, влияющих на продажи продукта

Описание задачи: Выявить влияние демографических и рыночных факторов на продажи продукции компании. Датасет marketing_campaigns.csv, содержащий рекламные кампании и соответствующие им продажи.

Задания:

Загрузите и предварительно обработайте данные.

Выполните регрессионный анализ влияния затрат на рекламу и возраста целевой аудитории на итоговую прибыль.

Постройте график остатков и проверьте адекватность модели.

Задача №6. Генерация интерактивных дашбордов с помощью Dash

Описание задачи: Создать интерактивный дашборд для сравнения динамик продаж в регионах. Файл region_sales.csv, содержащий продажи по регионам за разные периоды.

Задания:

Прочитайте и подготовьте данные.

Разработайте интерактивный дашборд с фильтрацией по региону и диапазону дат.

Отразите динамику продаж в каждом регионе с возможностью сравнивать регионы.

Задача №7. Аналитика покупателей с помощью кластерного анализа

Описание задачи: Определить сегменты покупателей интернет-магазина, исходя из их покупок и предпочтений. Данные: customer_purchases.csv, содержащий истории заказов клиентов (ID покупателя, сумма заказа, частота покупок).

Задания:

Предподготовьте данные, удалите выбросы и заполните недостающую информацию.

Используйте метод k-means для разделения покупателей на сегменты.

Определите оптимальное количество кластеров и проанализируйте поведение каждого сегмента.

Задача №8. A/B-тестирование рекламных баннеров

Описание задачи: Оценить эффективность нового рекламного баннера по сравнению с существующим вариантом. Файлы banner_clicks_old.csv и banner_clicks_new.csv, содержащие клики по старым и новым баннерам соответственно.

Задания:

Загрузите данные и сравните показатели кликов старых и новых баннеров.

Проведите статистический тест (например, z-test) для проверки гипотезы о превосходстве нового баннера.

Сделайте вывод о целесообразности замены старого баннера на новый.

Задача №9. Анализ группового поведения покупателей

Описание задачи: Исследовать влияние социального статуса покупателей на частоту покупок, используя регрессионный анализ. Датасет social_status_and_purchases.csv, содержащий социальные статусы покупателей и историю покупок.

Задания:

Заполнить пробелы в данных, подготовить столбцы для анализа.

Используя бинарную логистическую регрессию, выясните зависимость частоты покупок от социального статуса.

Интерпретируйте коэффициенты модели и дайте заключение о влиянии социальных статусов на покупки.

Задача №10. Визуализация качественных данных с помощью круговых диаграмм

Описание задачи: Постройте качественную круговую диаграмму, отражающую долю участников опроса по различным категориям (пол, возраст, профессия). Данные опроса респондентов содержатся в файле survey_results.csv.

Задания:

Импортируйте и очистите данные.

Составьте круговую диаграмму доли респондентов по выбранному признаку

(например, профессии).

Добавьте подписи и легенду для улучшения восприятия диаграммы.

Тема 4 «Машинное обучение и методы классификации и кластеризации»

Задача №1. Модели обучения с учителем и классификация кредитных заявок

Описание задачи: Вам предоставлена база данных кредитных заявок банка. Ваша задача – разработать модель классификации, способную автоматически определять риск дефолта заемщика. Нужно протестировать и сравнить три модели: логистическую регрессию, случайный лес и градиентный бустинг.

Задача №2. Полу-контролируемое обучение на смешанном датасете

Описание задачи: Используя частично размеченный датасет медицинских обследований, Ваша задача – реализовать полу-контролируемую модель для прогнозирования диагноза пациентов.

Задача №3. Градиентный бустинг с гиперпараметризацией

Описание задачи: Настройте гиперпараметры градиентного бустинга для достижения максимальной точности на соревновании Kaggle. Используйте GridSearchCV для нахождения лучших параметров.

Задача №4. K-Means кластеризация клиентов интернет-магазина

Описание задачи: Примените кластерный анализ для сегментирования клиентов интернет-магазина по их покупательскому поведению.

Задача №5. Деревья решений для диагностики заболевания

Описание задачи: Используя медицинские данные, нужно создать дерево решений для автоматической диагностики диабета у пациента.

Задача №6. Случайный лес для прогнозирования срока службы оборудования

Описание задачи: Используя инженерные данные, необходимо спрогнозировать ресурс оборудования с помощью модели случайного леса.

Задача №7. Кластеризация сотрудников по показателям продуктивности

Описание задачи: Используя данные о продуктивности сотрудников, произведите кластеризацию сотрудников по их производительности с помощью иерархического кластерирования.

Задача №8. Классификация email-писем на спам и не-спам

Описание задачи: Создайте классификатор для разделения входящих email-писем на спам и легитимные письма.

Задача №9. Самостоятельная настройка hyperparameters для алгоритма случайного леса

Описание задачи: Используя ручной подбор параметров, увеличьте точность случайного леса, экспериментируя с гиперпараметрами.

Задача №10. Нейронные сети для классификации снимков болезней кожи

Описание задачи: Реализуйте нейронную сеть на TensorFlow/Keras для классификации фотографий кожных заболеваний.

Тема 5 «Глубокое обучение и нейронные сети»

Задача №1. Сверточная нейронная сеть (CNN) для распознавания рукописных цифр MNIST

Описание задачи: Классическая задача классификации изображений — распознавание рукописных цифр на датасете MNIST. Создайте CNN-модель с несколькими слоями свертки и плотными слоями для обучения и тестирования.

Задача №2. Рекуррентная нейронная сеть (RNN/LSTM) для прогнозирования временных рядов

Описание задачи: Создание RNN-модели для прогнозирования временного ряда, например, курсов валют или стоимости акций. Модель должна научиться экстраполировать будущее значение на основе предыдущих данных.

Задача №3. Трансформеры для перевода коротких предложений (NLP)

Описание задачи: Реализуйте Transformer-для простого двуязычного перевода, например, английские короткие предложения в русские эквиваленты.

Задача №4. Autoencoder для сжатия изображений

Описание задачи: Создайте автоэнкодер для сжатия изображений, сохраняя ключевые черты картинки. Затем восстановите изображение обратно и оцените потерю качества.

Задача №5. Генеративная состязательная сеть (GAN) для синтеза изображений лиц

Описание задачи: Генерируйте реалистичные лица, обучив генеративную состязательную сеть (GAN) на датасете CelebA. В результате ваша модель должна уметь синтезировать новые уникальные лица.

Задача №6. Свертка с батч-нормализацией и остаточными связями (ResNet)

Описание задачи: Реализуйте архитектуру ResNet для классификации изображений Fashion-MNIST. Цель — добиться высокого уровня точности, используя остатки и батч-нормализацию.

Задача №7. Deep Q-Learning для игры Flappy Bird

Описание задачи: Создайте нейросеть с использованием DQN (Deep Q-learning Network) для обучения агента играть в игру Flappy Bird.

Задача №8. Siamese Neural Networks для сопоставления похожих картинок

Описание задачи: Создайте сиамскую нейронную сеть для сопоставления схожих изображений, основываясь на евклидовом расстоянии между изображениями.

Задача №9. Word Embeddings и Fine-Tuning с моделью Bert для NLP-задачи

Описание задачи: Используя pretrained-трансформер (BERT), улучшите модель классификации отзывов пользователей. Например, положительные отзывы vs негативные.

Задача №10. UNet для сегментации изображений клеток крови

Описание задачи: Создайте U-Net архитектуру нейронной сети для сегментации микроскопических изображений клеток крови. Данная задача требует выделения границ эритроцитов и лейкоцитов.

Тема 6 «Распределённые вычисления и параллельные алгоритмы»

Задача №1. MapReduce для подсчета частот встречаемости слов

Описание задачи: Используя парадигму MapReduce, посчитайте частоту появления каждого слова в большом тексте. Важно учесть большие объемы данных и необходимость распараллеливания вычислений.

Задача №2. Apache Spark для обработки больших лог-файлов

Описание задачи: Используя Apache Spark, напишите программу для анализа большого файла журналов веб-серверов. Программа должна агрегировать запросы по IP-адресам и выводить топ-10 самых активных IP-адресов.

Задача №3. Apache Spark для анализа временных серий

Описание задачи: Напишите приложение на Apache Spark для расчёта средних значений температурных измерений за каждый месяц на протяжении нескольких лет. Входные данные содержат миллионы записей с датой и температурой.

Задача №4. HDFS и чтение больших файлов с сохранением промежуточных результатов

Описание задачи: Запись данных в HDFS и последующий расчёт суммы чисел в огромных файлах с использованием параллельного чтения и обработки блоков данных.

Задача №5. Система массового обнаружения аномалий с использованием Spark Streaming

Описание задачи: Создайте систему с использованием Apache Spark Streaming для обнаружения редких событий в потоке данных в реальном времени.

Задача №6. Apache Spark для агрегации и сортировки по частоте

Описание задачи: Агрегируйте большое количество уникальных ключей и отсортируйте их по количеству вхождений. Требуется высокая производительность благодаря параллельному выполнению.

Задача №7. Анализ временных серий с Apache Spark SQL

Описание задачи: Используя Apache Spark SQL, рассчитайте максимальные суточные температуры по городу за последний год.

Задача №8. Обработка больших данных с Apache Flink

Описание задачи: Использование Apache Flink для обработки потоков данных в реальном времени. Организуйте вычисление суммарных транзакций по клиентам.

Задача №9. Kafka для потоковой передачи данных в реальном времени

Описание задачи: Реализуйте передачу данных через Kafka с отправкой сообщений и последующей обработкой потребителем.

Задача №10. Распределённое обучение нейронной сети с использованием Apache Spark

Описание задачи: Обучите небольшую нейронную сеть на больших данных с использованием распределённого обучения на Apache Spark.

Тема 7 «Интеграция и преобразование данных»

Задача №1. Автоматическая чистка данных с идентификацией аномалий

Описание задачи: Создать скрипт для автоматического удаления дубликатов и некорректных записей из датасета о погоде. В датасете имеются поля: температура, влажность, давление, скорость ветра. Необходимо очистить данные и сохранить их в формате .parquet.

Задача №2. Многопоточная интеграция данных из нескольких источников

Описание задачи: Организовать интеграцию данных из двух разных источников: Twitter API и GitHub API. Написать многопоточный скрипт для объединения данных и сохранения их в единой таблице.

Задача №3. Трансформация и объединение структурированных данных (CSV/XML/JSON)

Описание задачи: Собрать и объединить данные из нескольких источников разного формата (CSV, XML, JSON) в единый датафрейм. Данные имеют общую сущность (клиенты), но хранятся отдельно.

Задача №4. Чистота данных и выявление ошибок ввода

Описание задачи: Создать скрипты для идентификации и устранения неверных или противоречащих друг другу данных в датасете пользователей. Данные содержат атрибуты: ФИО, адрес, телефон, email.

Задача №5. Трансформация и перенос данных из PostgreSQL в MongoDB

Описание задачи: Необходима программа для миграции данных из реляционной базы данных (PostgreSQL) в документную базу данных (MongoDB). База данных содержит таблицу с заказами клиентов.

Задача №6. Валидация и обогащение данных (EDA)

Описание задачи: Написать сценарий для проведения первичного анализа данных (Exploratory Data Analysis) и обогащения данных дополнительными признаками, такими как количество пустых значений и процент заполнения данных.

Задача №7. Передача данных через Kafka и Apache Spark

Описание задачи: Отправить поток данных через брокер Kafka и обработать их с помощью Apache Spark для агрегации и вычисления статистики.

Задача №8. Развёртывание конвейера ETL с Airflow

Описание задачи: Автоматизировать процесс ETL с помощью инструмента Apache Airflow для регулярного обновления данных в производственной среде.

Задача №9. Преобразование и чистота JSON-данных

Описание задачи: Создать утилиту для преобразования неструктурированных

JSON-данных в пригодный для анализа формат. Данные должны содержать имена, адреса и телефоны пользователей.

Задача №10. Разработка ETL-конвейера с использованием Prefect

Описание задачи: Создать рабочий ETL-конвейер с использованием инструмента Prefect для управления задачами и контроля выполнения. Дан конвейер должен загружать данные из базы данных MySQL, обрабатывать их и сохранять в новую таблицу.

Тема 8 «Анализ поведения пользователей и персонализация услуг»

Задача №1. Профилирование клиентов на основе анализа поведения

Описание задачи: Используя данные о действиях пользователей (просмотры страниц, покупки, лайки), разработать систему профилирования клиентов для определения склонности к покупке определенных категорий товаров.

Задача №2. Collaborative Filtering для рекомендации фильмов

Описание задачи: Создать рекомендательную систему фильмов на основе совместных предпочтений пользователей (collaborative filtering). Пользователям предлагаются фильмы, которые понравились другим пользователям с похожими вкусами.

Задача №3. Контекстные рекомендации на основе контента (Content-Based Recommender)

Описание задачи: Реализовать рекомендательную систему на основе описания товара (content-based recommendation system). Пользователи получают рекомендации товаров, аналогичных ранее просмотренным или купленным ими товарам.

Задача №4. Таргетированная реклама на основе поведения пользователей

Описание задачи: Разработать систему таргетированной рекламы, учитывающей прошлые просмотры и предпочтения пользователей. Рекомендуемые объявления показываются на основе недавних взаимодействий пользователя с контентом сайта.

Задача №5. A/B-тестирование рекламной кампании

Описание задачи: Провести A/B-тестирование двух вариантов рекламной кампании. Цель – выяснить, какая версия приносит больше конверсий (нажатий на объявление, переходов на сайт).

Задача №6. Персонализированные push-уведомления

Описание задачи: Реализовать систему отправки персональных уведомлений пользователям на основе их прошлых покупок и текущего поведения на сайте. Рекомендуется отправлять уведомления с товарами, подобными недавно просматриваемым.

Задача №7. Personalized Offers на основе кластеризации

Описание задачи: Создать систему персональных предложений, которая группирует пользователей на основе их покупок и предлагаете каждому сегменту индивидуализированные скидки и акции.

Задача №8. Optimized Product Recommendation Engine

Описание задачи: Создать рекомендательную систему, использующую гибридный подход (content-based + collaborative filtering), чтобы предлагать покупателям самые точные и привлекательные товары.

Задача №9. Оценка эффективности персонификации

Описание задачи: Провести оценку эффективности системы персонализации, проверив, насколько увеличилось количество успешных сделок и удовлетворённость пользователей после введения рекомендаций.

Задача №10. Улучшение Retention Rate через персонализацию

Описание задачи: Создать систему, направленную на увеличение удержания пользователей путём предоставления персональных скидок и специальных предложений на основе их прошлого поведения.

Тема 9 «Прогнозирование и принятие решений»

Задача №1. Линейная регрессия для прогнозирования расходов на электроэнергию

Описание задачи: Используя исторические данные о расходах электроэнергии, создать модель линейной регрессии для прогнозирования ежемесячных расходов на энергию. Данные содержат информацию о площади помещения, числе жильцов и температуре окружающей среды.

Задача №2. ARIMA для прогнозирования продаж в ритейле

Описание задачи: Построить модель ARIMA для прогнозирования недельных продаж в магазине на ближайший месяц. Исторические данные о продажах предоставлены за предыдущий год.

Задача №3. Оптимизация запасов с помощью регрессии

Описание задачи: Создать модель регрессии для прогнозирования спроса на товар в онлайн-магазине, чтобы оптимизировать запасы и избежать излишков. Имеются данные о прошлом спросе, цене товара и сезонности.

Задача №4. Долгосрочный прогноз цен на жилье с использованием LSTM

Описание задачи: Используя исторический ценовой ряд недвижимости, создать модель на основе Long Short Term Memory (LSTM) для прогнозирования цен на жильё на следующий год.

Задача №5. Прогнозирование фондового индекса с использованием Time Series Forecasting

Описание задачи: Создать модель прогнозирования индексов фондового рынка на ближайшую неделю, используя данные о котировках за последние полгода.

Задача №6. Оптимизация маршрута доставки с использованием Genetic Algorithm

Описание задачи: Оптимизировать маршруты доставки товаров по городам, учитывая расстояние и стоимость транспортировки. Реализовать генетический алгоритм для поиска оптимального маршрута.

Задача №7. Прогнозирование потребления электроэнергии с использованием Profit Optimization

Описание задачи: Создать модель для прогнозирования потребления электроэнергии домохозяйствами с целью оптимизации тарифов и снижения пиковых нагрузок.

Задача №8. Прогнозирование объема производства с использованием Seasonal Decomposition

Описание задачи: Используя сезонный анализ временных рядов, построить прогноз объема производства на предприятии с учетом циклических колебаний спроса.

Задача №9. Optimal Inventory Management с использованием Reinforcement Learning

Описание задачи: Применение reinforcement learning для оптимизации складских запасов и уменьшения потерь от дефицита или избытка товаров.

Задача №10. Оптимизация расписания рейсов с использованием Mixed Integer Programming

Описание задачи: Используя mixed integer programming, оптимизировать расписание авиарейсов, минимизируя задержки и затраты на обслуживание самолётов.