Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Гриб Владислав Валерьевич
Должность: Ректор
Дата подписания: 14.10.2026 17:08:07
Уникальный программный ключ:
637517d24e103c3db032acf37e839d98ec1c5bb2f5eb89c29abfcd7f43985447

Образовательное частное учреждение высшего образования
«МОСКОВСКИЙ УНИВЕРСИТЕТ ИМЕНИ А.С. ГРИБОЕДОВА»
(ИМПЭ им. А.С. Грибоедова)

МЕЖДУНАРОДНЫЙ ИНСТИТУТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И БИЗНЕС-ИНФОРМАТИКИ

УТВЕРЖДАЮ
Директор Международного института информационных технологий и бизнес-информатики
_____ Т.В. Новикова
«17» декабря 2025г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по учебной дисциплине:

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

направление подготовки
38.03.05
БИЗНЕС-ИНФОРМАТИКА
(уровень бакалавриат)

направления/профиль:
«Цифровая трансформация бизнеса»

Форма обучения: очная, очно-заочная

Москва

Фонд оценочных средств для дисциплины «Статистические методы анализа данных». Направление подготовки/специальность 38.03.05 Бизнес-информатика, направленность (профиль): «Цифровая трансформация бизнеса»/ Т.В. Новикова – М.: ИМПЭ им. А.С. Грибоедова.

Фонд оценочных средств является неотъемлемой частью рабочей программы дисциплины.

Разработчик: Т.В. Новикова



Заведующий кафедрой _____/Т.В. Новикова, к. э. н., доцент

1. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

Настоящий Фонд оценочных средств (ФОС) является неотъемлемым приложением к рабочей программе дисциплины (РПД) «Статистические методы анализа данных». На данный ФОС распространяются все реквизиты утверждения, представленные в РПД по данной дисциплине.

2. ПЕРЕЧЕНЬ ОЦЕНОЧНЫХ СРЕДСТВ

Для определения качества освоения обучающимися учебного материала по дисциплине используются следующие оценочные средства:

№ п/п	Оценочное средство	Краткая характеристика оценочного средства	Представление оценочного средства в ФОС
1	Тестирование	Вид контроля, позволяющий оценить изученный теоретический материал.	Вопросы для проведения тестирования
2	Практические задания	Вид контроля, позволяющий оценить умение обучающегося применять осваиваемую компетенцию в практических ситуациях и при решении производственных задач	Задания к практическому занятию
3	Контрольная работа	Вид контроля, позволяющий определить результат освоения компетенций по дисциплине в рамках рассматриваемой темы, оцениваемый с помощью соответствующих индикаторов достижения компетенций	Задания контрольной работы
4	Самостоятельная работа	Вид контроля, позволяющий оценить проработку теоретического материала, изучение рекомендуемой литературы, выполнение практико-ориентированных заданий (заполнение таблиц, проведение сравнительного анализа, составление схем и др.), решение практических задач, создание презентаций, написание рефератов, подборку нормативного и иного материала и выполнение других заданий	Задания самостоятельной работы
5	Курсовая работа	Вид контроля, позволяющий выявить степень владения базовыми знаниями, умениями и навыками, необходимыми для обучения, и определить уровень владения новым материалом	Индивидуальные задания (темы) для курсовой работы
6	Зачет/Зачет с оценкой/Экзамен	Вид контроля, позволяющий выявить степень овладения знаниями, умениями и навыками, необходимыми для дальнейшего освоения образовательной программы подготовки	Вопросы для подготовки к зачету/зачету с оценкой/экзамену

3. ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ ДИСЦИПЛИНЕ

3.1. Сопроводительная информация.

Разработчик	Т.В. Новикова
Кафедра	Бизнес-информатика технологий управления
Наименование дисциплины	Статистические методы анализа данных
Факультет / институт	Институт международной экономики, лидерства и менеджмента
Направление подготовки / специальность	38.03.05 Бизнес-информатика
Количество вопросов в оценочных заданиях (диапазон)	От 20 до 40
Общее время тестирования (мин)	90
Общее количество вопросов/заданий в ФОС	40
Размещенность на сайте Университета примерного перечня вопросов, заданий ФОС – для подготовки обучающихся к прохождению оценки (да / нет)	Да

3.2. Характеристика оцениваемых компетенций

Код компетенции	Формулировка компетенции	Индикаторы достижения компетенции
ПК-15.	**ПК-15.** Способен информировать заказчика на этапе предконтрактных работ о возможностях типовой ИС и вариантах ее модификации	**ИПК-15.1 Знать**: Сетевые протоколы Основы современных операционных систем Основы современных СУБД Устройство и функционирование современных ИС Архитектура мультиарендного программного обеспечения Основы ИБ организации Современные стандарты информационного взаимодействия систем Программные средства и платформы инфраструктуры информационных технологий организаций Современные подходы и стандарты автоматизации организации Основы теории систем и системного анализа **ИПК-15.2 Уметь**: Проводить презентации заинтересованным сторонам в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС **ИПК-15.3 Владеть**: Навыками информирования заказчика на этапе предконтрактных работ о возможностях типовой ИС и вариантах ее модификации

4. СОДЕРЖАНИЕ ОЦЕНОЧНЫХ СРЕДСТВ ТЕКУЩЕГО КОНТРОЛЯ

4.1. ТИПОВЫЕ ТЕСТОВЫЕ ЗАДАНИЯ.

Тесты содержат набор вопросов, в полном объеме охватывающие изученный теоретический материал по указанной теме. Выполнение тестов позволяет определить результат освоения компетенций по дисциплине в рамках рассматриваемой темы, оцениваемый с помощью соответствующих индикаторов достижения компетенций. Индивидуальный тестовый сеанс для каждого обучающегося формируется по специальному алгоритму, обеспечивающему заданную тематическую структуру и пропорциональное наличие вопросов разного типа и сложности.

При формировании тестов необходимо использовать задания следующих типов:

Тип задания 1. Задания закрытого типа на установление соответствия.

Тип задания 2. Задания закрытого типа на установление последовательности.

Тип задания 3. Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных

Тип задания 4. Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных

Тип задания 5. Задания открытого типа с развернутым ответом.

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td>Раздел 1.
Тема 1.1 Введение в статистические методы анализа данных.</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания закрытого типа на установление соответствия.
Вариант 1. Установите соответствие между статистическим понятием и его определением.<table><tr><td>Понятие</td><td>Определение</td></tr><tr><td>А. Генеральная совокупность</td><td>1. Подмножество объектов, отобранных из генеральной совокупности для проведения анализа.</td></tr><tr><td>Б. Выборка</td><td>2. Множество всех возможных объектов или наблюдений, относительно которых делается вывод в исследовании.</td></tr><tr><td>В. Статистический признак</td><td>3. Величина, значение которой зависит от случая и может быть описана с помощью вероятностного распределения.</td></tr><tr><td>Г. Случайная величина</td><td>4. Характеристика единицы наблюдения, которая подлежит измерению или регистрации в ходе исследования.</td></tr></table>Правильный ответ: А–2, Б–1, В–4, Г–3

Вариант 2. Установите соответствие между этапом статистического исследования и его содержанием.<table><tr><td>Этап исследования</td><td>Содержание этапа</td></tr><tr><td>А. Сбор данных</td><td>1. Применение методов описательной статистики, проверка гипотез, построение моделей.</td></tr><tr><td>Б. Представление данных</td><td>2. Формулировка выводов на основе полученных статистических</td></tr></table></td></tr>
</table>

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><td></td><td>результатов в контексте предметной области.</td></tr>
<tr><td>В. Анализ данных</td><td>3. Организация систематического получения информации об объектах исследования (опросы, эксперименты, наблюдения).</td></tr>
<tr><td>Г. Интерпретация результатов</td><td>4. Визуализация и табулирование данных для выявления основных закономерностей (графики, таблицы, диаграммы).</td></tr>
</table>

Правильный ответ: А–3, Б–4, В–1, Г–2</td></tr>
<tr><td>Раздел 1.
Тема 1.2 Методы первичной обработки и представления данных.</td><td>ПК-15</td><td>ИПК-15.1</td><td>Задания закрытого типа на установление последовательности.
Вариант 1. Установите правильную последовательность этапов первичной обработки статистических данных.
1. Расчет сводных статистических показателей (среднее, медиана, мода).
2. Сбор «сырых» данных в результате наблюдения или эксперимента.
3. Проверка и очистка данных от ошибок и пропусков.
4. Группировка и упорядочивание данных (например, построение ранжированного ряда).
Правильный ответ: 2 → 3 → 4 → 1

Вариант 2. Установите правильную последовательность действий при подготовке данных для их последующего представления и анализа.
1. Выбор подходящего метода визуализации (гистограмма, диаграмма, график).
2. Кодирование качественных признаков для машинной обработки.
3. Формирование окончательной таблицы данных (датафрейма) со всеми переменными.
4. Проверка полноты и достоверности собранной информации.
Правильный ответ: 4 → 2 → 3 → 1</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
Раздел 1. Тема 1.3 Методы выявления и устранения аномальных наблюдений.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** При проектировании базы данных для хранения результатов социологического опроса вы обнаружили, что в поле «Возраст респондента» встречаются значения: 25, 34, 42, 150, 28, –5. Какой из перечисленных методов наиболее целесообразно использовать на этапе проектирования для предотвращения подобных аномалий в будущем? A) Удаление всех строк с аномальными значениями после загрузки данных. B) Применение правила трёх сигм (3σ) ко всем числовым полям. C) Введение ограничений (constraints) на диапазон допустимых значений при создании таблицы (например, CHECK (age BETWEEN 0 AND 120)). D) Замена аномальных значений средним по выборке. Правильный ответ: C Обоснование: При проектировании структуры базы данных важно заложить механизмы предотвращения некорректных данных на у СУБД, а не только их последующей очистки. Введение ограничений (CHECK-ограничений) — это стандартное средство проектирования реляционных баз данных, которое обеспечивает целостность данных на этапе ввода. Остальные методы (A, B, D) — это постфактумные процедуры очистки, которые не решают проблему на у проектирования структуры. **Вариант 2.** При разработке программного модуля для автоматической загрузки данных в аналитическую систему возникла необходимость фильтрации аномальных значений в количественных переменных. Какой из перечисленных подходов наиболее эффективно сочетает простоту реализации и адаптивность к различным распределениям данных? A) Удаление всех значений, выходящих за пределы интервала [Q1 – 1.5·IQR; Q3 + 1.5·IQR]. B) Удаление всех значений, превышающих среднее более чем на 2 стандартных отклонения.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			C) Ручная проверка каждого наблюдения экспертами. D) Замена всех экстремальных значений на медиану. Правильный ответ: A Обоснование: Метод межквартильного размаха (IQR) не предполагает нормальности распределения и устойчив к наличию самих выбросов, в отличие от методов, основанных на среднем и стандартном отклонении (вариант B). Это делает его универсальным и надёжным инструментом для первичной автоматизированной очистки данных в программных системах. Его легко реализовать в коде, и он хорошо интегрируется в конвейеры обработки данных, что важно при проектировании структур и логики работы с данными. Варианты C и D либо не масштабируются, либо искажают данные без необходимости.
Раздел 2. Тема 2.1 Корреляционно-регрессионный анализ. Модель множественной регрессии.	ПК-15	ПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных** **Вариант 1.** При построении модели множественной линейной регрессии для анализа факторов, влияющих на производительность труда сотрудников, исследователь получил высокий коэффициент детерминации (R2=0,92), но при этом некоторые регрессионные коэффициенты оказались статистически незначимыми. Какие из перечисленных действий целесообразно предпринять для улучшения модели? Выберите все верные варианты: A) Удалить из модели незначимые предикторы и перестроить регрессию. B) Увеличить объём выборки, чтобы повысить мощность статистических тестов. C) Заменить все незначимые коэффициенты на нули без пересчёта модели. D) Проверить наличие мультиколлинеарности между независимыми переменными. E) Игнорировать статистическую значимость, так как высокий R^2 гарантирует качество модели. Правильные ответы: A, B, D

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Обоснование: A: Удаление незначимых предикторов (особенно при подозрении на мультиколлинеарность) — стандартный шаг в построении интерпретируемой и устойчивой модели. B: Небольшой объём выборки может привести к низкой мощности тестов, из-за чего даже реальные эффекты не будут признаны значимыми. Увеличение выборки — корректный путь. D: Высокая корреляция между предикторами (мультиколлинеарность) завышает стандартные ошибки коэффициентов, делая их незначимыми, несмотря на реальное влияние. Это необходимо проверить (например, через VIF). Почему не C и E: C нарушает логику регрессионного анализа: модель должна быть пересчитана после изменения состава переменных. E — ошибка: высокий R^2 может быть следствием переобучения или мультиколлинеарности и не гарантирует ни адекватности, ни прогностической способности модели. **Вариант 2.** Какие из перечисленных условий (предпосылок) должны выполняться для корректного применения классической модели множественной линейной регрессии и валидной интерпретации её результатов? Выберите все верные варианты: A) Линейная зависимость между зависимой переменной и каждой из независимых переменных. B) Нормальное распределение всех независимых переменных. C) Гомоскедастичность (постоянство дисперсии) случайных ошибок. D) Отсутствие автокорреляции остатков. E) Равенство нулю среднего значения зависимой переменной. Правильные ответы: A, C, D

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Обоснование: Классическая линейная регрессия опирается на ряд ключевых предпосылок (Гаусса–Маркова): A: Модель предполагает линейную связь в параметрах. C: Гомоскедастичность необходима для эффективности и несмещённости оценок МНК. D: Отсутствие автокорреляции остатков (особенно важно для временных рядов) обеспечивает корректность стандартных ошибок. Почему не B и E: B: Независимые переменные могут иметь любое распределение; нормальность требуется только для ошибок (для построения доверительных интервалов и тестов). E: Среднее зависимой переменной не обязано быть нулём; это требование относится к математическому ожиданию случайной ошибки (E[ε] = 0), а не к Y.
Раздел 2. Тема 2.2 Фиктивные переменные. Критерий Чоу проверки однородности.	ПК-15	ПК-15.1	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Экономист строит модель зависимости заработной платы сотрудника от его стажа работы. В компании работают как мужчины, так и женщины. Чтобы учесть возможное различие в оплате труда между полами, он решил использовать фиктивную переменную. Вопрос: 1. Что такое фиктивная переменная и как её можно задать в этой задаче? 2. Как будет выглядеть уравнение регрессии с учётом этой переменной? 3. Почему важно учитывать такие качественные признаки (например, пол) при построении экономических моделей? Развернутый ответ: 1. Фиктивная переменная — это числовая переменная (обычно 0 или 1), которая кодирует качественный признак. Например: o D=1, если сотрудник — мужчина, o D=0, если сотрудник — женщина. 2. Уравнение регрессии:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Зарплата=$\beta_0+\beta_1\cdot$Стаж$+\beta_2\cdot D+\varepsilon$ Здесь β_2 покажет, на сколько в среднем отличается зарплата мужчин от зарплаты женщин при одинаковом стаже. 3. Если не учитывать такие факторы, модель может давать неточные или несправедливые выводы. Учёт качественных признаков делает модель более реалистичной и полезной для анализа и управления — это часть системного подхода к моделированию. **Вариант 2.** Аналитик подозревает, что зависимость между расходами на обучение персонала и производительностью труда до пандемии и после пандемии была разной. Он хочет проверить, изменилась ли структура этой зависимости. Вопрос: 1. Как называется статистический тест, который позволяет сравнить две регрессионные модели до и после определённого события? 2. Что означает, если этот тест показал статистически значимый результат? 3. Почему важно проверять такие гипотезы при анализе экономических процессов? Развернутый ответ: 1. Это критерий Чоу (тест на структурный сдвиг или однородность). 2. Если результат значим, это означает, что зависимость между переменными действительно изменилась — то есть одна общая модель для всего периода не подходит. Нужно строить отдельные модели для «до» и «после». 3. Экономические и организационные процессы часто меняются под влиянием внешних факторов (кризисы, реформы, технологии). Проверка однородности помогает избежать ошибок в прогнозах и рекомендациях, делая модели более адекватными реальности — это ключевой элемент математического моделирования
Раздел 2.	ПК-15	ПК-15.1	**Задания закрытого типа на установление соответствия.**

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td>Тема 2.3
Множественные коэффициенты корреляции.</td><td></td><td></td><td>
Вариант 1. Установите соответствие между статистическим показателем и его описанием.
<table>
<tr><th>Показатель</th><th>Описание</th></tr>
<tr><td>А. Парный коэффициент корреляции</td><td>1. Характеризует долю дисперсии зависимой переменной, объяснённую совместным влиянием всех независимых переменных в модели.</td></tr>
<tr><td>Б. Множественный коэффициент корреляции (R)</td><td>2. Показывает силу линейной связи между двумя переменными без учёта влияния других факторов.</td></tr>
<tr><td>В. Коэффициент детерминации (R^2)</td><td>3. Измеряет тесноту линейной связи между зависимой переменной и совокупностью независимых переменных.</td></tr>
<tr><td>Г. Частный коэффициент корреляции</td><td>4. Отражает силу связи между двумя переменными при исключении влияния других учтённых факторов.</td></tr>
</table>
Правильный ответ: А–2, Б–3, В–1, Г–4

Вариант 2. Установите соответствие между характеристикой множественной регрессии и её свойством.
<table>
<tr><th>Характеристика</th><th>Свойство</th></tr>
<tr><td>А. Множественный коэффициент корреляции (R)</td><td>1. Всегда принимает значения от 0 до 1.</td></tr>
<tr><td>Б. Коэффициент детерминации (R^2)</td><td>2. Может использоваться для сравнения качества разных моделей (с осторожностью).</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><td>В. Скорректированный R^2</td><td>3. Учитывает количество независимых переменных и размер выборки.</td></tr>
<tr><td>Г. Парный коэффициент корреляции</td><td>4. Может быть положительным или отрицательным.</td></tr>
</table>
Правильный ответ: А–1, Б–2, В–3, Г–4
</td></tr>
<tr><td>Раздел 2.
Тема 2.4 Статистический анализ адекватности модели множественной регрессии.</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания закрытого типа на установление последовательности.
Вариант 1. Установите правильную последовательность этапов статистического анализа адекватности модели множественной регрессии.
1. Проверка статистической значимости отдельных коэффициентов регрессии с помощью t-теста.
2. Оценка общей значимости модели с помощью F-теста.
3. Анализ остатков на гомоскедастичность, нормальность и автокорреляцию.
4. Расчёт коэффициента детерминации R^2 и скорректированного R^2_{adj}.

Правильный ответ: 4 → 2 → 1 → 3

Вариант 2. Установите правильную последовательность действий при интерпретации результатов множественной регрессии.
1. Проверить, значим ли F-критерий модели в целом.
2. Убедиться, что остатки модели не нарушают предпосылок (например, нет гетероскедастичности).
3. Проанализировать значения коэффициентов регрессии и их экономический смысл.
4. Оценить, насколько хорошо модель объясняет зависимую переменную (по R_2).
Правильный ответ: 4 → 1 → 3 → 2</td></tr>
<tr><td>Раздел 2.
Тема 2.5 Мультиколлинеарность факторов.</td><td>ПК-15</td><td>ИПК-15.1</td><td>Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных
Вариант 1. При разработке программного модуля для прогнозирования спроса на товар вы построили регрессионную модель с двумя предикторами:</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			«цена товара» и «расходы на рекламу». Однако при анализе оказалось, что коэффициент корреляции между этими двумя переменными равен 0,92. Какая проблема может возникнуть при использовании такой модели в программном обеспечении? Варианты ответов: A) Модель будет давать неточные прогнозы из-за высокой дисперсии оценок коэффициентов. B) Программа не сможет сохранить данные в базе из-за нарушения нормальных форм. C) Время выполнения запроса к базе данных значительно увеличится. D) Все коэффициенты модели автоматически станут нулевыми. Правильный ответ: A Обоснование: Высокая корреляция между независимыми переменными (мультиколлинеарность) не искажает прогнозы, но сильно увеличивает стандартные ошибки коэффициентов, делая их нестабильными и трудно интерпретируемыми. Это снижает надёжность модели, особенно если она используется для анализа влияния отдельных факторов. При проектировании аналитического ПО важно учитывать такие статистические аспекты на этапе подготовки данных и выбора признаков. Остальные варианты не связаны с мультиколлинеарностью: B — проблема проектирования БД, C — проблема производительности, D — неверно. **Вариант 2.** При проектировании структуры данных для аналитической системы вы планируете включить в регрессионную модель следующие переменные: «доход клиента», «средний чек», «ежемесячные траты» и «остаток на счёте». Что следует сделать **до** загрузки этих данных в модель, чтобы избежать проблем с качеством анализа? Варианты ответов: A) Удалить все текстовые поля из таблицы. B) Проверить парные корреляции между этими переменными и при необходимости исключить или объединить сильно коррелирующие.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			C) Обязательно нормализовать все переменные. D) Создать отдельную таблицу для каждой переменной. Правильный ответ: B Обоснование: Перечисленные переменные могут быть сильными линейными функциями друг друга (например, «доход» и «траты» часто связаны), что вызывает мультиколлинеарность. На этапе проектирования аналитической системы важно предусмотреть предварительный анализ взаимосвязей между признаками. Это часть подготовки данных и проектирования логики обработки. Нормализация (C) не решает проблему мультиколлинеарности, а создание отдельных таблиц (D) — избыточно и не связано с этой задачей.
Раздел 3. Тема 3.1 Обобщенная линейная модель множественной регрессии. Обобщенный МНК	ПК-15	ПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных** **Вариант 1.** Аналитик строит регрессионную модель, но при диагностике остатков обнаруживает, что дисперсия ошибок не постоянна (присутствует гетероскедастичность). Какие из перечисленных утверждений верны в этой ситуации? Выберите все верные варианты: A) Оценки коэффициентов, полученные обычным МНК, остаются несмещёнными, но становятся неэффективными. B) Стандартные ошибки коэффициентов, рассчитанные обычным способом, могут быть некорректными, что делает t- и F-тесты ненадёжными. C) Необходимо немедленно удалить половину наблюдений из выборки. D) Можно применить обобщённый метод наименьших квадратов (ОМНК) для получения более надёжных оценок. E) Гетероскедастичность не влияет на качество модели, если R^2 высокий. Правильные ответы: A, B, D Обоснование: A — верно: при гетероскедастичности МНК-оценки несмещённы, но не имеют минимальной дисперсии (неэффективны). B — верно: стандартные ошибки завышаются или занижаются, поэтому проверка гипотез становится некорректной.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			D — верно: ОМНК (или робастные стандартные ошибки) — один из способов коррекции проблемы. Почему не C и E: C — произвольное и необоснованное действие, которое ухудшит качество модели. E — заблуждение: высокий R^2 не гарантирует адекватность модели; нарушение предпосылок МНК подрывает статистическую достоверность выводов. **Вариант 2.** Какие из перечисленных ситуаций являются основаниями для использования обобщённого МНК (ОМНК) вместо обычного МНК? Выберите все верные варианты: A) Наличие автокорреляции остатков (например, в данных временных рядов). B) Наличие гетероскедастичности (неравной дисперсии ошибок). C) Все независимые переменные принимают только целые значения. D) Зависимая переменная измерена в рублях. E) Известна структура ковариационной матрицы случайных ошибок (например, из теоретических соображений или предварительного анализа). Правильные ответы: A, B, E Обоснование: ОМНК применяется, когда ковариационная матрица ошибок не является скалярной (т.е. нарушены условия гомоскедастичности и/или отсутствия автокорреляции): A и B — классические случаи нарушения предпосылок МНК. E — необходимое условие для практического применения ОМНК: нужно знать (или оценить) структуру матрицы Ω. Почему не C и D: Тип шкалы переменных (целые, денежные единицы) не влияет на выбор между МНК и ОМНК. Решающим фактором является поведение остатков, а не формат данных.
Раздел 3.	ПК-15	ИПК-15.1	**Задания закрытого типа на установление последовательности.**

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
Тема 3.2 Гетероскедастичность. Методы выявления гетероскедастичности.			**Вариант 1.** Установите правильную последовательность действий при проверке регрессионной модели на наличие гетероскедастичности. 1. Построить график зависимости остатков от предсказанных значений (ŷ). 2. Рассчитать остатки модели (разницу между фактическими и предсказанными значениями). 3. Применить формальный статистический тест (например, тест Бреуша–Пагана или Уайта). 4. Сделать вывод о наличии или отсутствии гетероскедастичности. Правильный ответ: 2 → 1 → 3 → 4 **Вариант 2.** Установите правильную последовательность этапов обработки данных в программном модуле, если после построения регрессионной модели обнаружена гетероскедастичность. 1. Преобразовать переменные (например, взять логарифм зависимой переменной). 2. Подтвердить наличие гетероскедастичности с помощью графического анализа и статистического теста. 3. Перестроить модель с использованием робастных стандартных ошибок или ОМНК. 4. Проверить новую модель на выполнение предпосылок (включая гомоскедастичность). Правильный ответ: 2 → 1 → 3 → 4
Раздел 3. Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** При разработке модуля для анализа данных о доходах населения вы обнаружили, что дисперсия ошибок в регрессионной модели растёт с увеличением уровня дохода (гетероскедастичность). Какой из перечисленных подходов наиболее целесообразен для коррекции этой проблемы на этапе проектирования программного обеспечения? Варианты ответов:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			A) Удалить все наблюдения с высоким доходом. B) Использовать взвешенный метод наименьших квадратов (ВМНК), присваивая меньший вес наблюдениям с высоким доходом. C) Заменить все значения дохода их средним арифметическим. D) Игнорировать проблему, так как она не влияет на прогнозы. Правильный ответ: B Обоснование: Гетероскедастичность означает, что наблюдения с высоким доходом имеют большую дисперсию ошибки, то есть они менее точны. ВМНК решает эту проблему, уменьшая вес таких «зашумленных» наблюдений при оценке коэффициентов. Это позволяет получить более надёжные и эффективные оценки. Такой подход можно заложить в логику аналитического модуля ПО, и знанию методов обработки данных. Остальные варианты либо искажают данные (A, C), либо игнорируют статистическую проблему (D). **Вариант 2.** Аналитик строит модель зависимости расходов домохозяйства от его дохода. Он замечает, что разброс расходов выше у богатых семей. Чтобы учесть это при оценивании модели, он решает применить взвешенный МНК. Какой принцип следует использовать для выбора весов? Варианты ответов: A) Присвоить всем наблюдениям одинаковый вес, равный 1. B) Присвоить вес, обратно пропорциональный доходу домохозяйства. C) Присвоить вес, прямо пропорциональный доходу домохозяйства. D) Выбрать веса случайным образом. Правильный ответ: B Обоснование: Если дисперсия ошибки растёт с доходом, то дисперсия $\sigma \frac{2}{i} \propto$ Доход$_i$. В взвешенном МНК оптимальные веса выбираются как $w_i = \frac{1}{\sigma \frac{2}{i}}$, то есть обратно пропорционально дисперсии. Следовательно, вес должен быть обратно пропорционален доходу. Это обеспечивает большее влияние на оценку тех наблюдений, которые измерены точнее (с меньшей дисперсией).

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Понимание этого принципа важно при проектировании алгоритмов обработки данных в аналитических системах. Варианты A, C и D не учитывают структуру гетероскедастичности и приведут к неэффективным оценкам.
Раздел 3. Тема 3.4 Регрессионные модели с автокоррелированными остатками.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных** **Вариант 1.** При анализе временного ряда (например, ежемесячных продаж компании) вы построили регрессионную модель, но обнаружили положительную автокорреляцию остатков (ошибки в соседние месяцы похожи). Какие из перечисленных утверждений верны в этой ситуации? Выберите все верные варианты: A) Оценки коэффициентов регрессии, полученные обычным МНК, остаются несмещёнными. B) Стандартные ошибки коэффициентов, рассчитанные обычным МНК, будут занижены, что приведёт к завышению t-статистик. C) Прогнозы модели будут абсолютно точными, так как автокорреляция не влияет на них. D) Модель может давать ложные выводы о значимости факторов. E) Автокорреляция означает, что в данных отсутствует тренд. Правильные ответы: A, B, D Обоснование: A — верно: при автокорреляции МНК-оценки несмещённы, но неэффективны. B — верно: автокорреляция искажает оценку дисперсии ошибок, обычно занижая стандартные ошибки, из-за чего t-статистики кажутся выше, чем есть на самом деле. D — верно: из-за искажённых стандартных ошибок можно ошибочно признать незначимый фактор значимым. Почему не C и E:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			C — неверно: автокорреляция указывает на то, что модель не уловила часть закономерности (например, инерцию процесса), поэтому прогнозы могут быть систематически смещены. E — неверно: автокорреляция часто сопровождает тренды и циклы, а не исключает их. **Вариант 2.** Какие из перечисленных методов или действий могут помочь устранить или скорректировать проблему автокорреляции остатков в регрессионной модели? Выберите все верные варианты: A) Включение в модель лаговой (предыдущей) переменной зависимой величины (например, продажи в предыдущем месяце). B) Применение обобщённого метода наименьших квадратов (ОМНК) с учётом структуры автокорреляции (например, AR(1)). C) Удаление половины наблюдений из выборки. D) Использование робастных стандартных ошибок по Ньюи–Уэсту. E) Преобразование всех переменных в логарифмы. Правильные ответы: A, B, D Обоснование: A — верно: если автокорреляция вызвана инерцией процесса, добавление лага часто устраняет проблему. B — верно: ОМНК позволяет явно учесть структуру ковариационной матрицы ошибок (например, при AR(1)-процессе). D — верно: робастные ошибки Ньюи–Уэста корректируют стандартные ошибки в присутствии автокорреляции и гетероскедастичности, делая тесты надёжными. Почему не C и E: C — произвольное и контрпродуктивное действие, которое ухудшит качество модели. E — логарифмирование может помочь при гетероскедастичности, но не решает проблему автокорреляции.

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td>Раздел 3.
Тема 3.5 Устранение автокорреляции</td><td>ПК-15</td><td>ИПК-15.1</td><td>
Задания закрытого типа на установление соответствия.
Вариант 1. Установите соответствие между методом устранения автокорреляции и его кратким описанием.
<table>
<tr><th>Метод</th><th>Описание</th></tr>
<tr><td>А. Метод Кокрана–Оркатта</td><td>1. Корректирует стандартные ошибки, не изменяя оценки коэффициентов; подходит для временных рядов с автокорреляцией и гетероскедастичностью.</td></tr>
<tr><td>Б. Включение лаговой зависимой переменной</td><td>2. Итеративная процедура, в которой модель перестраивается с использованием преобразованных переменных на основе оценки коэффициента автокорреляции.</td></tr>
<tr><td>В. Робастные ошибки Ньюи–Уэста</td><td>3. Добавление в модель значения зависимой переменной за предыдущий период для учёта инерции процесса.</td></tr>
<tr><td>Г. Обобщённый МНК (ОМНК)</td><td>4. Прямое применение МНК к модели, в которой ошибки имеют известную ковариационную структуру (например, AR(1)).</td></tr>
</table>
Правильный ответ: А–2, Б–3, В–1, Г–4

Вариант 2. Установите соответствие между проблемой в регрессионной модели и подходящим способом её решения.
<table>
<tr><th>Проблема</th><th>Способ решения</th></tr>
<tr><td>А. Остатки коррелируют во времени (автокорреляция)</td><td>1. Применить взвешенный МНК с весами, обратными дисперсии.</td></tr>
<tr><td>Б. Дисперсия остатков растёт с ростом предсказанных значений</td><td>2. Использовать робастные стандартные ошибки или ОМНК с учётом AR(1)-структуры.</td></tr>
<tr><td>В. Высокая корреляция между независимыми переменными</td><td>3. Исключить одну из сильно коррелирующих переменных или использовать регуляризацию.</td></tr>
</table>
</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Г. Нелинейная связь между переменными — 4. Ввести в модель квадратичный или логарифмический член. Правильный ответ: А–2, Б–1, В–3, Г–4
Раздел 4. Тема 4.1 Кластерный анализ. Основные этапы кластерного анализа.	ПК-15	ПК-15.1	**Задания закрытого типа на установление последовательности.** **Вариант 1.** Установите правильную последовательность основных этапов кластерного анализа. 1. Интерпретация и профилирование полученных кластеров. 2. Выбор метода кластеризации (например, k-средних, иерархический). 3. Подготовка данных (стандартизация, кодирование, удаление пропусков). 4. Определение числа кластеров (например, по методу «локтя» или силуэту). 5. Сбор и выбор признаков для анализа. Правильный ответ: 5 → 3 → 2 → 4 → 1 **Вариант 2.** Установите правильную последовательность действий при применении метода k-средних. 1. Назначение каждого объекта ближайшему центроиду. 2. Вычисление новых центроидов как средних значений признаков в каждом кластере. 3. Задание числа кластеров k и случайный выбор начальных центроидов. 4. Повторение шагов до тех пор, пока центроиды не перестанут существенно меняться. Правильный ответ: 3 → 1 → 2 → 4
Раздел 4. Тема 4.2 Метрики расстояний. Свойства кластеров. Методы кластерного анализа.	ПК-15	ПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** Аналитик проводит сегментацию клиентов по двум признакам: «средний чек» (в рублях) и «частота покупок» (в разах за месяц). Оба признака — количественные, но измерены в разных единицах и имеют разные масштабы. Какую метрику расстояния наиболее целесообразно использовать для кластеризации? Варианты ответов:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			A) Евклидово расстояние — без предварительной обработки данных. B) Манхэттенское расстояние — без предварительной обработки данных. C) Евклидово расстояние — после стандартизации признаков. D) Расстояние Хэмминга. Правильный ответ: C Обоснование: Евклидово расстояние чувствительно к масштабу переменных. Если не стандартизировать данные, признак с большим размахом (например, «средний чек» в тысячах рублей) будет доминировать над признаком с малым размахом («частота покупок» — 1–10 раз), что исказит результаты кластеризации. Стандартизация (приведение к среднему 0 и стандартному отклонению 1) устраняет этот эффект. Расстояние Хэмминга (D) применяется только для качественных (категориальных) данных, поэтому не подходит. Таким образом, C — единственный корректный и практически обоснованный выбор, соответствующий принципам системного анализа и математического моделирования **Вариант 2.** При анализе опроса клиентов используются три качественных признака: «пол» (муж/жен), «регион» (Центр/Юг/Север), «тип клиента» (новый/постоянный). Какой метод кластерного анализа и метрика наиболее подходят для таких данных? Варианты ответов: A) Метод k-средних с евклидовым расстоянием. B) Иерархическая кластеризация с расстоянием Хэмминга или Жаккара. C) Метод k-средних с манхэттенским расстоянием. D) Алгоритм DBSCAN с евклидовым расстоянием. Правильный ответ: B Обоснование: Метод k-средних (A, C) и DBSCAN (D) работают с количественными данными и требуют вычисления «среднего» или плотности в метрическом пространстве, что некорректно для категориальных признаков. Для данных, состоящих только из категорий, следует

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			использовать иерархическую кластеризацию с метриками, предназначенными для таких данных: Расстояние Хэмминга — число несовпадающих признаков, Коэффициент Жаккара — мера схожести множеств. Это обеспечивает содержательную и математически корректную группировку
Раздел 4. Тема 4.3 Расстояния между кластерами. Методы группировки кластеров.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных** **Вариант 1.** При реализации иерархической кластеризации в аналитическом модуле программного обеспечения разработчик должен выбрать метод определения расстояния между кластерами. Какие из перечисленных методов корректно определяют расстояние между двумя кластерами? Выберите все верные варианты: A) Минимальное расстояние между любыми двумя объектами из разных кластеров (метод ближайшего соседа). B) Максимальное расстояние между любыми двумя объектами из разных кластеров (метод дальнего соседа). C) Расстояние между центрами (центроидами) кластеров. D) Среднее арифметическое всех попарных расстояний между объектами двух кластеров (метод средней связи). E) Расстояние между самыми часто встречающимися значениями признаков в кластерах. Правильные ответы: A, B, C, D Обоснование: Эти четыре метода — стандартные подходы в иерархической кластеризации: A — одиночная связь (single linkage), B — полная связь (complete linkage), C — центроидный метод, D — средняя связь (average linkage). Все они корректно определяют межкластерное расстояние на основе расстояний между объектами или их агрегатов.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Почему не E: «Самые часто встречающиеся значения» (мода) не определяют положение кластера в пространстве признаков и не позволяют вычислить метрическое расстояние. Этот подход не используется в классических алгоритмах кластеризации. Знание этих методов необходимо при проектировании гибких аналитических систем, где пользователь должен иметь возможность выбирать стратегию группировки. **Вариант 2.** Какие из перечисленных утверждений о методах группировки кластеров в иерархической кластеризации являются верными? Выберите все верные варианты: A) Метод «ближайшего соседа» (одиночная связь) склонен образовывать вытянутые, цепочкообразные кластеры. B) Метод «дальнего соседа» (полная связь) стремится формировать компактные, шарообразные кластеры примерно одинакового размера. C) Все методы группировки всегда дают одинаковое количество кластеров при одном и том же пороге объединения. D) Результат кластеризации может существенно зависеть от выбранного метода определения расстояния между кластерами. E) Метод средней связи менее чувствителен к выбросам, чем метод ближайшего соседа. Правильные ответы: A, B, D, E Обоснование: A — верно: одиночная связь легко «связывает» кластеры через близкие точки, даже если основные массы удалены. B — верно: полная связь требует, чтобы все объекты были близки, что приводит к компактности. D — верно: выбор метода напрямую влияет на форму и структуру дендрограммы.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Е — верно: усреднение сглаживает влияние отдельных экстремальных расстояний (выбросов). Почему не С: Разные методы дают разные дендрограммы, поэтому при фиксированном пороге расстояния число кластеров будет разным. Это знание критически важно при проектировании ПО для анализа данных (ПК-5), так как разработчик должен предусмотреть параметризацию метода и информировать пользователя о его влиянии на результат.
Раздел 4. Тема 4.4 Иерархические методы кластеризации. Агломеративные методы.	ПК-15	ИПК-15.1	**Задания закрытого типа на установление последовательности.** **Вариант 1.** Установите правильную последовательность шагов агломеративного иерархического метода кластеризации. 1. Объединить две ближайшие группы (кластера) в один. 2. Рассчитать матрицу расстояний между всеми парами объектов. 3. Повторять объединение до тех пор, пока все объекты не окажутся в одном кластере. 4. Каждый объект считается отдельным кластером. Правильный ответ: 4 → 2 → 1 → 3 **Вариант 2.** Установите правильную последовательность действий при построении дендрограммы с использованием агломеративного метода. 1. Выбрать метод определения расстояния между кластерами (например, «ближайший сосед»). 2. Начать с n*n* кластеров, где n*n* — число объектов. 3. Обновить матрицу расстояний после каждого объединения. 4. Отобразить результат в виде древовидной диаграммы (дендрограммы). Правильный ответ: 1 → 2 → 3 → 4
Раздел 4. Тема 4.5 Иерархические дивизимные методы кластеризации.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** Аналитик разрабатывает модуль для сегментации клиентов крупной компании. Исходная выборка содержит 10 000 клиентов. Какой из перечисленных подходов наиболее точно описывает идею дивизимного (разделительного) иерархического метода кластеризации?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Варианты ответов: A) Каждый клиент сначала считается отдельным кластером, затем постепенно объединяются самые близкие. B) Все клиенты сначала считаются одним кластером, который последовательно делится на более мелкие группы. C) Клиенты случайным образом распределяются по заранее заданному числу кластеров. D) Клиенты группируются только на основе одного признака — например, региона проживания. Правильный ответ: B Обоснование: Дивизимный (от лат. *divisio* — разделение) метод начинается с одного кластера, включающего все объекты, и затем рекурсивно делит его на подкластеры, выбирая на каждом шаге наиболее «неоднородную» группу для разбиения. Это противоположно агломеративному подходу (вариант A). Варианты C и D описывают соответственно k-средние и простую сортировку, а не иерархическую дивизимную кластеризацию. **Вариант 2.** При реализации дивизимного метода кластеризации в программном обеспечении разработчик должен выбрать алгоритм для разбиения кластера на две части. Какой из перечисленных методов чаще всего используется на этапе деления в дивизимных подходах? Варианты ответов: A) Метод ближайшего соседа. B) Алгоритм k-средних с k=2. C) Метод главных компонент. D) Линейная регрессия. Правильный ответ: B Обоснование: На практике дивизимные методы часто используют k-средние с k=2 для разбиения текущего кластера на две подгруппы, так как это эффективный и устойчивый способ выявить внутреннюю структуру. Хотя чисто дивизимные алгоритмы (например, DIANA) используют другие эвристики, в современных программных реализациях (в том числе в

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			библиотеках Python, R) k-средние — популярный инструмент для этапа деления. Методы A, C и D не предназначены для прямого разбиения кластера на две однородные группы: A — метод агломеративной кластеризации, C — метод снижения размерности, D — метод регрессионного анализа.
Раздел 4. Тема 4.6 Примеры реализации иерархических методов кластеризации.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных** **Вариант 1.** При разработке модуля иерархической кластеризации в аналитическом программном обеспечении разработчик должен учитывать особенности хранения и обработки данных. Какие из перечисленных утверждений о реализации иерархических методов верны? Выберите все верные варианты: A) Агломеративные методы требуют хранения полной матрицы попарных расстояний между объектами, что может быть ресурсоёмко при больших объёмах данных. B) Дивизимные методы обычно более эффективны по памяти, так как не требуют хранения всей матрицы расстояний сразу. C) Результат иерархической кластеризации удобно представлять в виде дендрограммы, для которой требуется древовидная структура данных. D) Иерархические методы всегда быстрее, чем метод k-средних, независимо от размера выборки. E) При большом числе объектов (>10 000) агломеративные методы могут стать вычислительно непрактичными без оптимизации. Правильные ответы: A, C, E Обоснование: A — верно: матрица расстояний имеет размер n×n, что при n=10 000 даёт 100 млн элементов — серьёзная нагрузка на память. C — верно: дендрограмма — это дерево, и для её хранения и визуализации используются древовидные структуры данных (например, списки или графы).

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			E — верно: сложность агломеративного метода — $O(n^3)$ или $O(n^2 \log n)$, что делает его медленным на больших данных. Почему не B и D: B — неверно: дивизимные методы также часто требуют расчёта расстояний внутри кластеров, и их эффективность зависит от реализации; они не обязательно экономнее по памяти. D — неверно: k-средние обычно значительно быстрее ($O(n)$–$O(n^2)$), особенно на больших данных. **Вариант 2.** При проектировании программного модуля для иерархической кластеризации разработчик должен учитывать особенности алгоритмов и структуры данных. Какие из перечисленных утверждений о практической реализации иерархических методов являются верными? Выберите все верные варианты: A) Агломеративный метод можно эффективно реализовать с использованием очереди с приоритетом (priority queue) для хранения пар кластеров по возрастанию расстояния. B) Для хранения истории объединений кластеров (дендрограммы) удобно использовать структуру данных «дерево» или список связей (linkage matrix). C) Дивизимный метод DIANA реализуется путём последовательного применения алгоритма k-средних с k=2 ко всем кластерам на каждом у . D) Вычислительная сложность наивной реализации агломеративной кластеризации — $O(n^3)$, что требует оптимизации при работе с большими данными. E) Иерархические методы не требуют предварительной стандартизации данных, так как автоматически учитывают масштаб признаков. Правильные ответы: A, B, D Обоснование: A — верно: очередь с приоритетом (например, куча) позволяет быстро находить пару кластеров с минимальным расстоянием, что ускоряет алгоритм.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			B — верно: дендрограмма — это дерево, и в программных реализациях (например, в SciPy) она хранится как матрица связей (linkage matrix), где каждая строка описывает одно объединение. D — верно: наивный алгоритм требует n−1n−1 итераций, на каждой из которых пересчитывается матрица расстояний размером $O(n^2)$, что даёт общую сложность $O(n^3)$. Почему не C и E: C — неверно: классический дивизимный алгоритм DIANA использует эвристику на основе дисперсии и наиболее рассеянного объекта, а не k-средние. Хотя k-средние *можно* использовать для деления, это не определение DIANA. E — неверно: иерархические методы чувствительны к масштабу признаков, как и большинство метрических методов. Без стандартизации признак с большим диапазоном будет доминировать.
Раздел 4. Тема 4.7 Итеративные методы. Метод к-средних.	ПК-15	ПК-15.1	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Метод k-средних используется для группировки похожих объектов. 1. Как алгоритм определяет, к какому кластеру отнести объект? 2. Почему важно стандартизировать (привести к одному масштабу) признаки перед применением этого метода? 3. Приведите один пример из жизни, где можно использовать метод k-средних (например, в магазине, банке, школе). Развёрнутый ответ: 1. Объект относят к тому кластеру, центр которого ближе всего к этому объекту (по расстоянию). 2. Если признаки измеряются в разных единицах (например, возраст — в годах, а доход — в тысячах рублей), то признак с большими числами будет «весить» больше. Стандартизация делает все признаки равноправными. 3. Например, в банке можно сгруппировать клиентов по уровню дохода и сумме вклада, чтобы предложить разные банковские продукты разным группам.

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>Вариант 2. Представьте, что вы анализируете данные о студентах: их успеваемость (средний балл) и участие в олимпиадах (количество побед). Вы хотите разделить их на группы с помощью метода k-средних.
1. Что будет центром кластера в этом случае?
2. Что может произойти, если выбрать слишком маленькое число кластеров (например, k=1k=1)?
3. Почему этот метод полезен для анализа данных в образовании или бизнесе?

Развёрнутый ответ:
1. Центр кластера — это среднее значение по успеваемости и среднее количество побед в олимпиадах для всех студентов в этой группе.
2. Если k=1k=1, то все студенты окажутся в одной группе, и мы не увидим различий между ними — кластеризация не даст полезной информации.
3. Метод помогает найти скрытые группы (например, талантливых, средних, отстающих) и принимать решения: кому давать стипендию, кого приглашать на курсы и т.д. Это делает управление более обоснованным и эффективным.</td></tr>
<tr><td>Раздел 4. Тема 4.8 Метод к-средних. Критерии качества кластеризации.</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания закрытого типа на установление соответствия.
Вариант 1. Установите соответствие между названием критерия качества кластеризации и его кратким описанием.
<table>
<tr><td>Критерий</td><td>Описание</td></tr>
<tr><td>А. Сумма квадратов внутри кластеров (WCSS)</td><td>1. Чем выше значение, тем лучше разделены кластеры.</td></tr>
<tr><td>Б. Метод «локтя»</td><td>2. Графический способ выбора числа кластеров
k по точке излома на графике WCSS.</td></tr>
<tr><td>В. Коэффициент силуэта</td><td>3. Общая мера компактности и раздельности кластеров; принимает значения от –1 до +1.</td></tr>
</table></td></tr>
</table>

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><td>Г. Межкластерная дисперсия</td><td>4. Сумма квадратов расстояний от объектов до центроидов их кластеров; стремится к минимуму.</td></tr>
</table>
Правильный ответ: А–4, Б–2, В–3, Г–1

Вариант 2. Установите соответствие между целью анализа и подходящим критерием для оценки результата кластеризации методом k-средних.
<table>
<tr><td>Цель</td><td>Критерий</td></tr>
<tr><td>А. Выбрать оптимальное число кластеров k</td><td>1. Сумма квадратов внутри кластеров (WCSS)</td></tr>
<tr><td>Б. Оценить, насколько плотно объекты лежат внутри своего кластера</td><td>2. Коэффициент силуэта</td></tr>
<tr><td>В. Проверить, насколько хорошо кластеры разделены друг от друга</td><td>3. Метод «локтя» (на основе WCSS)</td></tr>
<tr><td>Г. Получить общую оценку качества всей кластеризации</td><td>4. Коэффициент силуэта</td></tr>
</table>
Правильный ответ: А–3, Б–1, В–4, Г–4
</td></tr>
<tr><td>Раздел 4.
Тема 4.9 Метод нечетких к-средних</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных
Вариант 1. Чем отличается метод нечётких k-средних (Fuzzy C-Means) от классического метода k-средних?
Варианты ответов:
A) В нечётком методе каждый объект принадлежит только одному кластеру, как и в классическом.
B) В нечётком методе объект может частично принадлежать нескольким кластерам одновременно, с разной степенью принадлежности (от 0 до 1).
C) Нечёткий метод работает только с качественными данными, а классический — только с количественными.
D) Нечёткий метод всегда даёт больше кластеров, чем классический.</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Правильный ответ: B Обоснование: Ключевая идея нечёткой кластеризации — мягкое (soft) разбиение: вместо жёсткого отнесения объекта к одному кластеру, он получает степени принадлежности ко всем кластерам (например, 70% — кластер A, 30% — кластер B). Это особенно полезно, когда границы между группами размыты (например, клиенты со «смешанным» поведением). Такой подход соответствует концепции нечётких вычислений и позволяет строить более гибкие модели реальных экономических и социальных процессов. Остальные варианты неверны: A описывает классический метод, C и D — заблуждения. **Вариант 2.** В каком из перечисленных случаев целесообразнее использовать метод нечётких k-средних, а не классический? Варианты ответов: A) При сегментации клиентов, где каждый клиент чётко относится к одному типу (например, «только пенсионер» или «только студент»). B) При анализе данных, где границы между группами не чёткие, и объекты могут обладать признаками нескольких кластеров одновременно. C) Когда все признаки измерены в бинарной шкале (0/1). D) Когда нужно получить самый быстрый результат без потери точности. Правильный ответ: B Обоснование: Метод нечётких k-средних специально разработан для ситуаций, где объекты не могут быть однозначно отнесены к одной группе. Например, клиент может быть «частично лояльным» и «частично ценовым». В таких случаях нечёткая модель даёт более реалистичное и информативное описание. Классический метод (A) подходит для чётких категорий, но не для размытых. Варианты C и D не являются преимуществами нечёткого подхода (он даже медленнее классического).
Раздел 5. Тема 5.1 Понятие методов классификации.	ПК-15	ПК-15.1	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Методы классификации широко используются в анализе данных. В чём главное отличие задачи классификации от задачи кластеризации?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Приведите один пример задачи классификации из области экономики, бизнеса или управления. Укажите, какие данные могут быть признаками, а что — целевой переменной (классом). Как результаты классификации могут помочь в принятии управленческого решения? Развернутый ответ: Отличие: В классификации классы (группы) известны заранее, и модель учится относить новые объекты к уже существующим категориям. В кластеризации группы неизвестны, и алгоритм сам находит структуру в данных. Пример: Банк хочет определить, выдаст ли клиент кредит (да/нет). Признаки: возраст, доход, кредитная история, наличие работы. Целевая переменная (класс): «одобрить» или «отказать». Польза для управления: Модель классификации помогает банку автоматизировать решение, снизить риск невозврата и ускорить обработку заявок. Это делает процесс более объективным, быстрым и экономически выгодным — что соответствует целям системного анализа и математического моделирования **Вариант 2.** Аналитик работает с двумя задачами: — Задача А: предсказать, купит ли клиент товар («да» или «нет»). — Задача Б: предсказать, на какую сумму купит клиент товар (в рублях). 1. Какая из этих задач — задача классификации, а какая — задача регрессии? 2. Назовите один метод, который можно использовать для решения задачи классификации. 3. Почему важно правильно определить тип задачи (классификация или регрессия) при разработке модели? Развернутый ответ: 1. Задача А — классификация, потому что целевая переменная категориальная (два класса: «да» / «нет»).

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Задача Б — регрессия, потому что целевая переменная числовая (сумма покупки в рублях). 2. Пример метода классификации: метод ближайших соседей (k-NN), логистическая регрессия, дерево решений и т.п. 3. Потому что разные типы задач требуют разных методов и метрик качества. Если применить регрессию к задаче классификации (или наоборот), модель будет работать некорректно, и её прогнозы нельзя будет использовать для принятия решений. Правильное определение типа задачи — основа надёжного математического моделирования.
Раздел 5. Тема 5.2 Логистическая регрессия.	ПК-15	ПК-15.1	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Логистическая регрессия — один из самых популярных методов классификации. 1. Для какого типа задач применяется логистическая регрессия? Приведите пример цели (что именно предсказывает модель). 2. Чем отличается логистическая регрессия от обычной (линейной) регрессии? 3. Приведите один пример использования логистической регрессии в экономике, бизнесе или управлении. Как это помогает принимать решения? Развернутый ответ: 1. Логистическая регрессия применяется для задач бинарной классификации, где нужно предсказать вероятность принадлежности объекта к одному из двух классов. Например: предсказать вероятность того, что клиент откажется от услуги («да» / «нет»). 2. Отличие: В линейной регрессии прогнозируется числовое значение (например, доход). В логистической регрессии прогнозируется вероятность принадлежности к классу (значение от 0 до 1), которое затем можно перевести в решение (например, «если вероятность > 0,5 — класс “да”»). 3. Пример: Банк использует логистическую регрессию, чтобы определить, выдаст ли заёмщик кредит. Модель анализирует возраст, доход, кредитную

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>историю и выдаёт вероятность дефолта. Если вероятность высока — кредит не дают. Это снижает финансовые риски и делает процесс принятия решений быстрее и объективнее — что соответствует принципам системного анализа и математического моделирования

Вариант 2. Представьте, что вы построили модель логистической регрессии для прогнозирования того, купит ли клиент новый товар («да» или «нет»). В модели использованы два признака:
— X_1: возраст клиента (в годах),
— X_2: количество покупок за последний месяц.
1. Что означает, если коэффициент при X_2 оказался положительным?
2. Почему модель выдаёт вероятность, а не сразу ответ «да» или «нет»?
3. Как менеджер может использовать эту вероятность на практике?

Развернутый ответ:
1. Положительный коэффициент при X_2 означает, что чем больше покупок клиент делал в прошлом месяце, тем выше вероятность, что он купит новый товар. То есть активность клиента положительно влияет на его склонность к покупке.
2. Модель выдаёт вероятность, потому что это даёт более гибкую и информативную оценку. Например, вероятность 0,9 — почти уверенность, а 0,55 — слабый сигнал. Это лучше, чем жёсткий «да/нет», особенно при неопределённости.
3. Менеджер может:
o Направить рекламу только тем клиентам, у кого вероятность покупки выше, например, 70%,
o Составить персональные предложения для групп с разной вероятностью,
o Оценить ожидаемый спрос и спланировать закупки.</td></tr>
<tr><td>Раздел 5. Тема 5.3 Метрики качества классификации.</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания закрытого типа на установление соответствия.
Вариант 1. Установите соответствие между метрикой качества бинарной классификации и её определением.
<table><tr><td>Метрика</td><td>Определение</td></tr></table></td></tr>
</table>

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Тест по теме</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><td>А. Точность (Accuracy)</td><td>1. Доля правильно предсказанных положительных объектов среди всех объектов, которые модель назвала положительными.</td></tr>
<tr><td>Б. Полнота (Recall)</td><td>2. Доля правильно предсказанных положительных объектов среди всех фактических положительных объектов.</td></tr>
<tr><td>В. Точность положительного класса (Precision)</td><td>3. Общая доля правильно классифицированных объектов (и положительных, и отрицательных).</td></tr>
<tr><td>Г. F1-мера</td><td>4. Среднее гармоническое между точностью и полнотой.</td></tr>
</table>

Правильный ответ: А–3, Б–2, В–1, Г–4
Вариант 2. Установите соответствие между ситуацией и наиболее подходящей метрикой качества.
<table>
<tr><th>Ситуация</th><th>Метрика</th></tr>
<tr><td>А. Важно не пропустить ни одного больного (например, при диагностике редкого заболевания)</td><td>А. Полнота (Recall)</td></tr>
<tr><td>Б. Важно, чтобы большинство «тревог» были действительно опасными (например, фрод-детекция в банке)</td><td>Б. Точность (Precision)</td></tr>
<tr><td>В. Классы сбалансированы, и нужно оценить общую эффективность модели</td><td>В. Точность (Accuracy)</td></tr>
<tr><td>Г. Нужно учесть и точность, и полноту в одной метрике</td><td>Г. F1-мера</td></tr>
</table>

Правильный ответ: А–1, Б–2, В–3, Г–4
</td></tr>
<tr><td>Раздел 5. Тема 5.4 Метрики качества</td><td>ПК-15</td><td>ИПК-15.1</td><td>Задания закрытого типа на установление последовательности.</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
классификации. Регуляризация в задаче регрессии.			**Вариант 1.** Установите правильную последовательность этапов при построении регрессионной модели с регуляризацией (например, гребневой регрессии или Lasso) в аналитическом программном модуле. 1. Разделить данные на обучающую и тестовую выборки. 2. Оценить качество модели на тестовой выборке с помощью метрик (например, MSE, MAE). 3. Обучить модель регрессии с регуляризацией на обучающей выборке. 4. Выбрать значение параметра регуляризации (например, с помощью кросс-валидации). 5. Подготовить и стандартизировать данные (особенно важно для регуляризации). Правильный ответ: 1 → 5 → 4 → 3 → 2 **Вариант 2.** Установите правильную последовательность действий при сравнении качества двух моделей классификации в программной системе. 1. Рассчитать матрицу ошибок (confusion matrix) для каждой модели. 2. Вычислить метрики качества: точность, полноту, F1-меру. 3. Применить обе модели к одной и той же тестовой выборке. 4. Сделать вывод, какая модель лучше подходит для поставленной задачи. 5. Убедиться, что тестовая выборка не участвовала в обучении моделей. Правильный ответ: 5 → 3 → 1 → 2 → 4
Раздел 5. Тема 5.5 Методы деревьев решений.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** При разработке модуля анализа клиентских данных в CRM-системе вы решили использовать дерево решений для прогнозирования оттока клиентов. Какое из перечисленных утверждений наиболее точно отражает преимущество дерева решений с точки зрения проектирования программного обеспечения? Варианты ответов: А) Дерево решений требует, чтобы все признаки были нормально

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			распределены. B) Дерево решений даёт интерпретируемую модель в виде правил «если–то», которую легко встроить в бизнес-логику системы. C) Дерево решений всегда даёт более точные прогнозы, чем другие методы. D) Дерево решений невозможно реализовать без использования нейронных сетей. Правильный ответ: B Обоснование: Деревья решений генерируют понятные человеку правила (например: «Если возраст > 50 И доход < 30 тыс., ТО вероятность оттока высока»). Это делает их легко интегрируемыми в экспертные системы, CRM, BI-платформы и другие ПО, где важна прозрачность принятия решений. Варианты A, C и D — неверны: деревья не требуют нормальности, не всегда точнее (например, уступают ансамблям), и реализуются без нейросетей. **Вариант 2.** Какая структура данных наиболее естественно подходит для хранения и обработки модели дерева решений в программном обеспечении? Варианты ответов: A) Линейный массив B) Связный список C) Дерево (древовидная структура) D) Хеш-таблица Правильный ответ: C Обоснование: Модель дерева решений по своей сути является древовидной структурой: узлы содержат условия, а рёбра — переходы к дочерним узлам или листьям (решениям). Для эффективной навигации, обхода и сериализации (сохранения/загрузки модели) в ПО используется именно древовидная структура данных.
Раздел 5. Тема 5.6 Ансамбли деревьев решений: Случайный лес.	ПК-15	ИПК-15.1	**Задания закрытого типа на установление последовательности.** **Вариант 1.** Установите правильную последовательность шагов при построении модели случайного леса.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			1. Объединить прогнозы всех деревьев (например, усреднением для регрессии или голосованием для классификации). 2. Для каждого дерева случайным образом выбрать подмножество обучающих объектов (с возвращением — бутстрэп). 3. Построить множество деревьев решений независимо друг от друга. 4. На каждом узле дерева случайным образом выбрать подмножество признаков для поиска лучшего разбиения. Правильный ответ: 2 → 4 → 3 → 1 **Вариант 2.** Установите правильную последовательность действий при применении обученного случайного леса к новому объекту в программной системе. 1. Передать новый объект каждому дереву в лесу. 2. Получить прогноз от каждого дерева. 3. Собрать все прогнозы и определить итоговый результат (например, по большинству голосов). 4. Убедиться, что признаки нового объекта закодированы и масштабированы так же, как в обучающей выборке. Правильный ответ: 4 → 1 → 2 → 3
Раздел 5. Тема 5.7 Ансамбли деревьев решений: Градиентный бустинг.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных** **Вариант 1.** Чем отличается градиентный бустинг от случайного леса при построении ансамбля деревьев решений? Варианты ответов: A) В градиентном бустинге все деревья строятся независимо друг от друга, как в случайном лесе. B) В градиентном бустинге каждое новое дерево обучается на ошибках предыдущих, последовательно улучшая модель. C) Градиентный бустинг использует только один признак для всех деревьев. D) Градиентный бустинг не может быть использован для задач регрессии.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			Правильный ответ: B Обоснование: Градиентный бустинг — это последовательный метод: каждое новое дерево пытается исправить ошибки, допущенные предыдущими. Это принципиальное отличие от случайного леса, где деревья строятся параллельно и независимо. Такой подход позволяет добиться высокой точности, но требует аккуратной настройки (например, скорости обучения). При проектировании аналитических модулей в ПО (ПК-5) важно понимать эту логику, чтобы правильно реализовать обучение, сохранение состояния модели и обработку данных (ИПК-5.1). Остальные варианты неверны: A описывает случайный лес, C и D — заблуждения. **Вариант 2.** При реализации градиентного бустинга в программном обеспечении разработчик должен учитывать особенности алгоритма. Какое из перечисленных утверждений наиболее точно отражает ключевую особенность градиентного бустинга? Варианты ответов: A) Модель состоит из большого числа слабых моделей (обычно простых деревьев), объединённых последовательно для минимизации функции потерь. B) Все деревья в ансамбле должны быть глубокими и сложными. C) Градиентный бустинг не требует разделения данных на обучающую и тестовую выборки. D) Алгоритм работает быстрее случайного леса при любом объёме данных. Правильный ответ: A Обоснование: Градиентный бустинг строит ансамбль последовательно, добавляя на каждом шаге дерево, которое уменьшает функцию потерь (например, MSE или логистические потери). Используются слабые модели (часто деревья глубины 1–3), что повышает обобщающую способность. Это знание необходимо при проектировании эффективных и надёжных модулей машинного обучения в ПО (ПК-5, ИПК-5.1). Варианты B, C и D неверны:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Тест по теме
			деревья обычно простые, разделение выборок обязательно, а скорость зависит от параметров и реализации.

Критерии оценивания тестового задания:

Оценка	Критерии оценивания
отлично	от 90 до 100 % правильно выполненных заданий
хорошо	от 70 до 89 % правильно выполненных заданий
удовлетворительно	от 50 до 69 % правильно выполненных заданий
неудовлетворительно	менее 50 % правильно выполненных заданий

4.2 ТИПОВЫЕ ПРАКТИЧЕСКИЕ ЗАДАНИЯ

Практические задания должны отражать умение обучающегося применять осваиваемую компетенцию в практических ситуациях и при решении производственных задач.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
Тема 1.2 Методы первичной обработки и представления данных.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** Вариант 1. Вы получили данные о заказах в сервисе доставки еды за один день: • номер заказа, • сумма заказа (в рублях), • район доставки, • тип клиента («новый», «постоянный»), • время доставки (в минутах). В данных обнаружены ошибки: • у некоторых заказов сумма = –800 руб., • в колонке «район» встречаются варианты: «Центр», «центр», «ЦЕНТР», • у нескольких заказов время доставки указано как 999 минут (очевидно, ошибка). Задание: 1. Как вы исправите или обработаете эти ошибки?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			2. Какой график вы построите, чтобы сравнить среднее время доставки в разных районах? 3. Как эта информация может помочь руководителю службы доставки? Развернутый ответ: 1. Отрицательные суммы — удалить или заменить на пропуск; все названия районов привести к единому виду (например, «Центр»); значения 999 мин — считать ошибкой и либо удалить, либо заменить на медиану по району. 2. Столбчатая диаграмма: по оси X — районы, по оси Y — среднее время доставки (в минутах). 3. Руководитель сможет выявить, в каких районах доставка занимает больше времени, и принять меры: добавить курьеров, скорректировать зоны ответственности или информировать клиентов о возможных задержках. **Вариант 2.** Интернет-магазин прислал данные о продажах за неделю: • дата, • сумма заказа, • город доставки, • статус заказа («выполнен», «отменён», «в пути»). Обнаружены проблемы: • один заказ имеет сумму –5000 руб., • в городе указано «Москва», «москва», «MOSCOW», • несколько заказов со статусом «неизвестно». Задание: 1. Что вы сделаете с ошибочными и несогласованными данными?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			2. Какой график поможет понять, в каких городах чаще отменяют заказы? 3. Как руководитель магазина может использовать этот график для улучшения работы? Развёрнутый ответ: 1. Отрицательную сумму — удалить или проверить; все названия городов привести к единому виду (например, «Москва»); статус «неизвестно» — либо уточнить, либо исключить из анализа отмен. 2. Столбчатая диаграмма: по оси X — города, по оси Y — доля отменённых заказов (%). 3. Руководитель сможет направить менеджеров в проблемные города, улучшить логистику или пересмотреть условия доставки.
Тема 1.3 Методы выявления и устранения аномальных наблюдений.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы анализируете данные о ежедневных продажах товара в розничной сети за последние 30 дней. В большинстве дней объём продаж колеблется от 80 до 120 единиц, но в один день зафиксировано **1500 единиц**. Такое значение может быть вызвано как ошибкой ввода данных, так и реальным событием (например, акцией или оптовой закупкой). Задание: 1. Какие простые методы вы можете использовать для проверки, является ли это значение аномалией? 2. Если вы установите, что это — ошибка, как вы её обработаете? 3. Почему важно не игнорировать такие наблюдения, даже если они кажутся «редкими»? Развёрнутый ответ:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Можно: 1. Построить ящик с усами (boxplot) — значение за пределами усов будет считаться выбросом; 2. Использовать правило межквартильного размаха (IQR): если значение > Q3 + 1.5·IQR — это аномалия; 3. Сравнить с историческими данными (были ли раньше подобные всплески?). - Если это ошибка — можно удалить запись или заменить на медиану (например, 100 ед.). Если есть основания полагать, что это реальное событие — оставить, но отметить отдельно. - Аномалии могут искажать средние значения, прогнозы и KPI. Например, из-за одного выброса средний спрос будет завышен, и склад закупит лишний товар. Поэтому их нужно либо корректировать, либо учитывать отдельно при планировании. **Вариант 2.** В системе мониторинга работы серверов фиксируется время отклика (в миллисекундах). Обычно оно составляет 50–200 мс, но иногда появляются значения 5000 мс и выше. Эти «всплески» могут указывать либо на технический сбой, либо на кратковременную перегрузку. Задание: 1. Почему такие аномальные значения важны для анализа, даже если они редки? 2. Какой метод выявления аномалий вы бы выбрали для автоматической диагностики в системе мониторинга и почему? 3. Как результаты этого анализа могут быть использованы для формализации процесса технической поддержки? Развёрнутый ответ:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			1. Даже редкие сбои влияют на надёжность и качество сервиса. Если их игнорировать, система может не успеть предотвратить крупный отказ. 2. Для автоматизации подойдёт метод IQR или пороговое правило (например, всё, что > 1000 мс — аномалия). Эти методы просты, быстро работают и не требуют сложных вычислений — что важно для систем мониторинга в реальном времени. 3. Можно настроить автоматическое оповещение: если за час зафиксировано более 3 аномалий — отправить сигнал дежурному инженеру. Это формализует реакцию на сбои и повышает надёжность ИТ-инфраструктуры
Тема 2.1 Корреляционно-регрессионный анализ. Модель множественной регрессии.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы участвуете в создании программы для магазина, которая будет предсказывать **количество** продаж товара за неделю. На основе прошлых данных построена модель множественной регрессии: Продажи=β_0+β_1·Цена+β_2·Расходы на рекламу+β_3·Наличие акции («Наличие акции» — 1, если акция есть, 0 — если нет.) Задание: 1. Какие таблицы и поля вы создадите в базе данных программы, чтобы хранить исходные данные, коэффициенты модели и результаты прогноза? 2. Как вы организуете ввод новых данных (например, цена и реклама на следующую неделю), чтобы они корректно обрабатывались программой? 3. Почему важно, чтобы программа могла сохранять несколько версий модели (например, после обновления данных)? Развёрнутый ответ:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			1. Таблица «Данные о продажах»: номер недели, цена, расходы на рекламу, наличие акции (да/нет), фактические продажи. Таблица «Модель прогноза»: номер модели, дата создания, коэффициенты (β_0, β_1, β_2, β_3), качество модели (например, R^2). Таблица «Прогнозы»: номер недели, предсказанные продажи, номер использованной модели. 2. В программе сделать форму ввода с проверкой: - цена — положительное число, - расходы на рекламу — неотрицательное число, - акция — только «да» или «нет». Некорректные значения не принимаются. 3. Если модель переобучают на новых данных, старая может оказаться точнее в некоторых ситуациях. Хранение нескольких версий позволяет сравнивать их и при необходимости вернуться к лучшей — это делает систему надёжнее **Вариант 2.** Сервисный центр хочет автоматически рассчитывать примерное время ремонта бытовой техники. Построена регрессионная модель: Время ремонта=β_0+β_1·Тип_техники+β_2 ·Сложность_поломки+β_3·Опыт_мастера Задание: 1. Как вы спроектируете структуру базы данных, чтобы система могла принимать заявки, применять модель и сохранять результаты? 2. Какие ограничения вы зададите в базе, чтобы избежать ошибок (например, отрицательного времени ремонта)? 3. Как мастер будет получать результат расчёта в повседневной работе?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Развёрнутый ответ: 1. Таблица «Заявки»: номер заявки, тип техники (холодильник, стиральная машина и т.д.), сложность поломки (1–5 баллов), опыт мастера (стаж в годах). Таблица «Модель расчёта»: коэффициенты модели, дата последнего обновления. Таблица «Результаты»: номер заявки, предсказанное время ремонта (в часах), дата расчёта. 2. В таблице «Заявки» установить правила: - сложность — целое число от 1 до 5, - опыт мастера — не меньше 0, - тип техники — только из заранее заданного списка. 3. После ввода данных о заявке программа автоматически рассчитывает время и показывает его в окне заявки. Это помогает мастеру планировать день, а клиенту — понимать, сколько ждать. Такая система требует чёткого проектирования данных и логики работы
Тема 2.3 Множественные коэффициенты корреляции.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для анализа эффективности работы отдела продаж. В модели используются три фактора: — расходы на рекламу, — число звонков менеджеров, — количество встреч с клиентами. Зависимая переменная — объём продаж за месяц. После расчётов получен множественный коэффициент корреляции R = 0,85. Задание: 1. Что показывает это значение? Достаточно ли его, чтобы считать модель хорошей?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			2. Какие таблицы и поля вы создадите в базе данных, чтобы хранить исходные данные, результаты расчётов и коэффициенты связи? 3. Почему важно хранить не только сам коэффициент R, но и коэффициент детерминации R^2? Развёрнутый ответ: 1. Значение R = 0,85 означает, что между объёмом продаж и совокупностью трёх факторов существует сильная линейная связь. Однако одного R недостаточно: нужно проверить значимость модели (например, по F-критерию) и убедиться, что нет переобучения. 2. Таблица «Данные по месяцам»: месяц, расходы на рекламу, число звонков, число встреч, объём продаж. Таблица «Результаты анализа»: дата расчёта, множественный коэффициент корреляции (R), коэффициент детерминации (R^2), F-статистика, p-значение. Таблица «Коэффициенты модели»: название фактора, значение коэффициента, стандартная ошибка. 3. Коэффициент R^2 показывает, какая доля изменчивости объёма продаж объясняется моделью (в данном случае — около 72%). Это ключевой показатель качества модели, который удобнее интерпретировать, чем R. Его хранение позволяет сравнивать разные версии модели и оценивать их полезность. **Вариант 2.** В учебном заведении анализируют успеваемость студентов. Зависимая переменная — средний балл. Факторы: — время, потраченное на самостоятельную работу, — посещаемость занятий,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			— участие в олимпиадах. Получен множественный коэффициент корреляции R = 0,62. Задание: Как вы спроектируете структуру базы данных, чтобы система могла регулярно обновлять модель и пересчитывать R при появлении новых данных? Какие ограничения вы зададите в таблицах, чтобы избежать ошибок (например, отрицательного времени на учёбу)? Как руководство может использовать значение R при принятии решений? Развёрнутый ответ: 1. Таблица «Студенты»: номер студента, курс, группа. Таблица «Учебные показатели»: номер студента, часы самостоятельной работы, посещаемость (%), участие в олимпиадах (да/нет), средний балл. Таблица «Статистика моделей»: дата расчёта, R, R^2, список использованных факторов. При добавлении новых данных запускается автоматический пересчёт коэффициентов. 2. В таблице «Учебные показатели» установить правила: - часы самостоятельной работы ≥ 0, - посещаемость от 0 до 100%, - средний балл от 2,0 до 5,0. 3. Значение R = 0,62 говорит о умеренной связи между факторами и успеваемостью. Это значит, что указанные показатели важны, но есть и другие влияющие причины. Руководство может: — усилить контроль за самостоятельной работой, — мотивировать участие в олимпиадах,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			— искать дополнительные факторы (например, условия проживания). Хранение R в структурированном виде позволяет отслеживать изменения во времени и оценивать эффект от управленческих решений.
Тема 2.5 Мультиколлинеарность факторов.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Экономист строит модель зависимости прибыли компании от трёх показателей: — расходы на рекламу, — число клиентов, — доход от продаж. При анализе выяснилось, что доход от продаж и число клиентов сильно связаны между собой: чем больше клиентов, тем выше доход. Это явление называется мультиколлинеарностью. Задание: Почему наличие мультиколлинеарности может исказить результаты модели, даже если она хорошо описывает данные? Как это может повлиять на принятие управленческого решения (например, стоит ли увеличивать расходы на рекламу)? Что можно сделать, чтобы устранить или уменьшить влияние мультиколлинеарности? Развёрнутый ответ: 1. При мультиколлинеарности коэффициенты модели становятся нестабильными: их значения могут сильно меняться при небольшом изменении данных. Также стандартные ошибки коэффициентов завышаются, из-за чего факторы могут казаться незначимыми, хотя на самом деле влияют на результат.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			2. Если модель покажет, что «доход от продаж» не значим (из-за коллинеарности с числом клиентов), руководство может ошибочно решить, что рост продаж не влияет на прибыль. Это приведёт к неверным выводам и, возможно, к сокращению активностей, которые на самом деле полезны. 3. Можно: — исключить один из сильно связанных факторов (например, оставить только «число клиентов»), — объединить их в один показатель (например, «средний чек» = доход / число клиентов), — использовать методы, устойчивые к коллинеарности (например, регуляризацию). Это повысит надёжность модели и качество принимаемых решений **Вариант 2.** Аналитик разрабатывает модель для оценки стоимости квартир. В неё включены следующие факторы: — общая площадь, — жилая площадь, — площадь кухни. Эти три показателя очень сильно коррелируют между собой: обычно, чем больше общая площадь, тем больше и жилая, и кухня. Задание: 1. Как мультиколлинеарность повлияет на интерпретацию коэффициентов модели? 2. Почему такая модель может быть опасной при автоматизации оценки стоимости (например, в онлайн-калькуляторе)? 3. Какие шаги вы предпримете как разработчик, чтобы обеспечить надёжность системы автоматической оценки?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Развёрнутый ответ: 1. Коэффициент при «жилой площади» может получиться отрицательным, хотя логично, что чем больше жилая площадь — тем дороже квартира. Это происходит потому, что модель «путает» вклад каждого фактора из-за их сильной связи. 2. Автоматическая система может начать давать странные оценки: например, предлагать более высокую цену за квартиру с меньшей жилой площадью. Это подорвёт доверие пользователей и приведёт к финансовым потерям. 3. Как разработчик: — проверить парные корреляции между факторами, — удалю или объединю сильно связанные переменные (например, оставлю только «общую площадь»), — протестировать модель на новых данных, — заложить в систему предупреждение, если входные данные нарушают логические связи (например, жилая площадь > общей). Это обеспечит устойчивость и достоверность автоматизированного решения
Тема 3.1 Обобщенная линейная модель множественной регрессии. Обобщенный МНК	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы анализируете данные о времени обработки заявок в технической поддержке компании. Зависимая переменная — время (в минутах). Факторы: — сложность заявки (баллы), — загруженность оператора (число активных заявок), — тип оборудования клиента. При проверке остатков выяснилось, что дисперсия ошибок растёт с увеличением времени: для сложных заявок разброс времени гораздо больше, чем для простых.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Это нарушение условия равной дисперсии (гомоскедастичности). Задание: 1. Почему в этой ситуации обычный метод наименьших квадратов (МНК) может дать ненадёжные результаты? 2. Какой метод следует использовать вместо обычного МНК и почему? 3. Как это повлияет на расчёт показателей эффективности работы службы поддержки? Развёрнутый ответ: 1. При неравной дисперсии (гетероскедастичности) оценки коэффициентов остаются несмещёнными, но их стандартные ошибки рассчитываются неверно. Из-за этого тесты на значимость факторов становятся недостоверными, и можно принять ложные выводы (например, признать важный фактор незначимым). 2. Следует использовать обобщённый метод наименьших квадратов (ОМНК). Он учитывает структуру дисперсии ошибок и даёт более точные и надёжные оценки, особенно когда дисперсия зависит от уровня зависимой переменной. 3. Более точная модель позволит правильно оценить влияние каждого фактора на время обработки. Например, можно выявить, что именно загруженность оператора, а не сложность заявки, — главная причина задержек. Это поможет обоснованно перераспределить нагрузку и точнее планировать ресурсы, что повысит объективность расчёта показателей эффективности **Вариант 2.** Компания прогнозирует расходы на обслуживание оборудования в зависимости от:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			— возраста оборудования (лет), — условий эксплуатации (шкала от 1 до 5), — количества предыдущих ремонтов. Анализ показал, что ошибки модели коррелируют во времени: если в одном месяце расходы были завышены, то и в следующем они тоже выше ожидаемого. Это указывает на автокорреляцию остатков. Задание: 1. Почему автокорреляция делает обычную регрессионную модель непригодной для надёжного прогноза? 2. Какой подход вы выберете для коррекции модели и почему? 3. Как использование обобщённой модели повлияет на формализацию процесса планирования бюджета на обслуживание? Развёрнутый ответ: 1. Автокорреляция означает, что в модели не учтена какая-то закономерность (например, сезонность или износ). Из-за этого прогнозы становятся систематически смещёнными, а доверительные интервалы — уже, чем должны быть. Это ведёт к ошибкам в планировании бюджета. 2. Воспользоваться обобщённым методом наименьших квадратов (ОМНК) с учётом структуры автокорреляции (например, предположу, что ошибки следуют модели авторегрессии первого порядка). Это позволит учесть зависимость между соседними наблюдениями и получить более точные оценки. 3. На основе скорректированной модели можно строить реалистичные прогнозы с корректными интервалами

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			неопределённости. Это позволит финансовой службе формализовать процесс бюджетирования: вместо фиксированной суммы закладывать диапазон, зависящий от состояния парка оборудования. Такой подход повышает надёжность и обоснованность расчётов.
Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для анализа расходов домохозяйств. Зависимая переменная — месячные расходы (в рублях). Факторы: доход, число членов семьи, город проживания. Анализ показал: у богатых семей разброс расходов гораздо больше, чем у бедных. Это означает, что дисперсия ошибок модели зависит от уровня дохода (гетероскедастичность). Задание: 1. Почему в этой ситуации обычная регрессия даст ненадёжные результаты? 2. Как вы измените структуру расчётов в программе, чтобы учесть эту особенность? 3. Какие таблицы и поля вы создадите в базе данных, чтобы поддерживать работу модели с весами? Развёрнутый ответ: 1. Обычная регрессия считает все наблюдения одинаково точными. Но у богатых семей данные «зашумлены» сильнее, поэтому их влияние на модель должно быть меньше. Без учёта этого модель будет смещённой и неустойчивой, особенно при прогнозировании для малообеспеченных семей. 2. В программе нужно реализовать взвешенный метод наименьших квадратов: — рассчитать вес для каждого домохозяйства как

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			обратную величину к дисперсии (например, вес = 1 / доход), — использовать эти веса при расчёте коэффициентов модели. 3. Таблица «Домохозяйства»: номер, доход, число членов, город, расходы. Таблица «Веса»: номер домохозяйства, вес (рассчитан по формуле). Таблица «Модель»: название коэффициента, значение, дата расчёта. Такая структура позволяет гибко обновлять веса и пересчитывать модель при поступлении новых данных. **Вариант 2.** Компания анализирует издержки на доставку товаров в разные регионы. Зависимая переменная — стоимость доставки. Факторы: расстояние, вес груза, тип дороги. Выяснилось, что в удалённых регионах разброс цен на доставку очень велик, а в центральных — стабилен. Это классический случай гетероскедастичности. Задание: 1. Как вы организуете ввод и хранение данных в базе, чтобы подготовить их к применению взвешенного МНК? 2. Какой механизм расчёта весов вы заложите в программу? 3. Как пользователь (логист) будет получать прогноз стоимости доставки с учётом этой модели? Развёрнутый ответ: 1. В таблице «Доставки» предусмотреть поля: регион, расстояние, вес, тип дороги, стоимость. Дополнительно создать справочник «Регионы» с признаком

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			«удалённость» (центр/периферия). Это позволит автоматически определять, для каких записей нужен больший вес. 2. Программа будет рассчитывать вес по формуле: — для центральных регионов: вес = 1, — для удалённых: вес = 0,3 (или другое значение, полученное из данных). Веса можно также рассчитывать как обратную величину к среднему разбросу по региону. 3. Логист вводит параметры заказа (расстояние, вес, регион). Программа автоматически определяет вес, применяет модель и выводит прогнозируемую стоимость. Результат отображается в окне расчёта заказа — это требует чёткого проектирования интерфейса и логики взаимодействия между данными и моделью
Тема 3.5 Устранение автокорреляции	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете систему для прогнозирования ежедневных продаж в магазине. В регрессионную модель входят: количество рекламных объявлений, погода и день недели. После построения модели выяснилось, что ошибки модели коррелируют во времени: если сегодня продажи выше прогноза, то завтра они тоже чаще выше. Это признак автокорреляции остатков. Задание: 1. Почему автокорреляция делает обычную регрессию непригодной для надёжного прогноза? 2. Как вы измените структуру хранения данных, чтобы учесть зависимость между соседними днями? 3. Какие дополнительные поля нужно добавить в таблицу с данными, чтобы реализовать метод устранения автокорреляции?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Развёрнутый ответ: 1. Автокорреляция означает, что в модели не учтена внутренняя динамика процесса (например, инерция спроса). Из-за этого прогнозы становятся систематически смещёнными, а оценки точности — оптимистичными. Это особенно опасно при планировании закупок. 2. Нужно хранить данные в строгом хронологическом порядке и обеспечить возможность обращения к предыдущим наблюдениям. Для этого в таблице обязательно должен быть порядковый номер дня или дата, по которой можно однозначно определить предыдущий и следующий период. 3. Следует добавить поле «продажи в предыдущий день» как новый фактор. Также можно создать отдельную таблицу «Остатки модели», где хранить ошибки за каждый день — это понадобится, если использовать методы, учитывающие структуру автокорреляции (например, преобразование Кохрана–Оркатта). **Вариант 2.** Компания анализирует ежемесячные расходы на электроэнергию в зависимости от: площади помещения, числа сотрудников и температуры на улице. Анализ остатков показал положительную автокорреляцию: высокие расходы в одном месяце часто сменяются высокими в следующем. Это связано с сезонностью и инерцией потребления. Задание: 1. Какой способ устранения автокорреляции вы выберете и почему? 2. Как вы спроектируете структуру базы данных, чтобы поддерживать этот метод?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			3. Почему важно предусмотреть такую структуру заранее, ещё на этапе проектирования системы учёта? Развёрнутый ответ: 1. Выбрать включение лаговой переменной — то есть добавить в модель расходы на электроэнергию за предыдущий месяц. Это просто, интерпретируемо и эффективно устраняет автокорреляцию, вызванную инерцией процесса. 2. В основной таблице «Энергопотребление» нужно предусмотреть: ○ поле «месяц» (с возможностью сортировки по возрастанию), ○ поле «расход_электроэнергии», ○ поле «расход_электроэнергии_предыдущий_месяц» (может заполняться автоматически при загрузке данных). Также, можно создать связь «один ко многим» между месяцами для удобного доступа к истории. 3. Если не заложить такую структуру заранее, потом будет сложно или невозможно добавить лаговые переменные без перестройки всей системы. Это приведёт к дополнительным затратам времени и ресурсов. Правильное проектирование структуры данных с учётом временной зависимости — ключ к надёжной и расширяемой системе.
Тема 4.3 Расстояния между кластерами. Методы группировки кластеров.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Компания проводит сегментацию клиентов по двум признакам: средний чек и частота покупок. Для этого используется иерархическая кластеризация. Возникает вопрос: как определять расстояние между уже образованными группами — по ближайшим объектам, по центрам или как-то ещё?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Задание: 1. Опишите два разных способа определения расстояния между кластерами (например, метод ближайшего и дальнего соседа). Чем они отличаются? 2. Как выбор метода повлияет на форму и количество полученных групп? 3. Почему важно заранее выбрать метод группировки при формализации процесса сегментации в информационной системе? Развёрнутый ответ: 1. Метод ближайшего соседа: расстояние между кластерами — это расстояние между самыми близкими объектами из разных групп. Метод дальнего соседа: расстояние — между самыми удалёнными объектами. Первый метод даёт вытянутые, цепочкообразные кластеры, второй — компактные и шарообразные. 2. При использовании метода ближайшего соседа может получиться мало крупных кластеров, даже если данные содержат мелкие обособленные группы. Метод дальнего соседа, наоборот, сохраняет компактность, но может дробить естественные скопления. 3. Если метод не зафиксирован заранее, то при каждом запуске система может давать разные результаты, что сделает сегментацию непредсказуемой. Для автоматизации маркетинговых решений (например, рассылки по группам) нужна устойчивая и воспроизводимая логика, поэтому выбор метода — часть формализации процесса **Вариант 2.** Аналитик строит иерархическую кластеризацию для группировки товаров по признакам:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			цена, популярность, сезонность. Руководство хочет использовать результаты для расчёта эффективности товарных групп и планирования закупок. Задание: 1. Почему выбор метода расчёта расстояния между кластерами влияет на инженерные расчёты показателей эффективности? 2. Какой метод вы порекомендуете — среднюю связь или центроидный метод — и почему? 3. Как вы обеспечите, чтобы результаты кластеризации можно было использовать в расчётах автоматически? Развёрнутый ответ: 1. Разные методы дают разные составы групп. Например, одна и та же группа товаров может включать разные позиции в зависимости от метода. Это напрямую влияет на расчёт средней рентабельности, оборачиваемости или спроса по группе — а значит, и на решения о закупках. 2. Я рекомендую метод средней связи, потому что он устойчив к выбросам и даёт сбалансированные кластеры. Центроидный метод может привести к инверсии (нарушению монотонности расстояний), что затруднит интерпретацию дендрограммы и автоматизацию. 3. Чтобы результаты можно было использовать автоматически, нужно: — сохранять номер кластера для каждого товара в справочной таблице, — фиксировать порог объединения (например, «объединять до расстояния 5»), — хранить правила группировки (метод, метрика) в настройках системы. Это позволит регулярно пересчитывать группы и

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			использовать их в других модулях (например, в модуле планирования закупок), что соответствует требованиям к формализации и расчёту показателей результативности
Тема 4.5 Иерархические дивизимные методы кластеризации.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для анализа ассортимента крупного магазина. В начале все товары считаются одной группой. Затем система должна последовательно делить эту группу на подгруппы (например, сначала — «продукты» и «непродукты», потом — «молочка», «мясо» и т.д.). Для этого используется дивизимный метод кластеризации. Задание: 1. Какие таблицы и поля вы создадите в базе данных, чтобы хранить результаты такого иерархического деления? 2. Как вы организуете логику выбора кластера для деления в программе? 3. Как пользователь (менеджер по ассортименту) будет просматривать и использовать полученную иерархию групп? Развёрнутый ответ: 1. Таблица «Товары»: номер товара, название, признаки (цена, вес, категория и т.д.). Таблица «Кластеры»: номер кластера, номер родительского кластера, уровень вложенности, описание. Таблица «Состав кластера»: номер товара, номер кластера. Такая структура позволяет строить дерево групп любой глубины. 2. Программа будет на каждом шаге: — рассчитывать меру неоднородности для каждого

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			текущего кластера (например, внутрикластерную дисперсию), — выбирать кластер с наибольшей неоднородностью для деления, — применять метод (например, k-средние с k=2) для разделения на две части. 3. Менеджер увидит древовидное меню в интерфейсе программы: корень — «Все товары», далее — подгруппы. Он сможет: — остановить деление на нужном у , — присвоить названия группам, — экспортировать состав каждой группы для анализа продаж. Это требует чёткого проектирования интерфейса и структуры данных**.** **Вариант 2.** Компания хочет сегментировать своих клиентов, начав с одного общего списка и последовательно разделяя его на всё более однородные подгруппы (например, сначала — по активности, затем — по доходу, затем — по интересам). Выбран дивизимный подход, так как он лучше отражает логику бизнес-анализа «от общего к частному». Задание: 1. Почему дивизимный метод удобнее для бизнес-аналитика, чем агломеративный, в этой ситуации? 2. Как вы спроектируете структуру базы данных, чтобы поддерживать процесс последовательного деления? 3. Какие ограничения вы введёте, чтобы избежать слишком мелких или пустых групп? Развёрнутый ответ:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			1. Дивизимный метод начинается с единой целостной картины (все клиенты), что соответствует реальному взгляду руководителя. Аналитик сам решает, по какому признаку важнее всего разделить клиентов на первом шаге, что делает процесс более управляемым и интерпретируемым. 2. Создадим таблицы: ○ «Клиенты»: номер, признаки (доход, активность и т.д.), ○ «Группы»: номер группы, номер родительской группы, уровень, правило деления (например, «доход > 100 тыс.»), ○ «Состав групп»: номер клиента, номер группы. Связь «группа → подгруппы» реализуется через поле «родительская группа». 3. В программе зададим правила: — минимальный размер группы (например, не менее 10 клиентов), — максимальная глубина дерева (например, не более 5 у й), — запрет на деление, если внутри группы все клиенты одинаковы по ключевым признакам. Это предотвратит бессмысленные разбиения и обеспечит устойчивость системы
Тема 4.7 Итеративные методы. Метод к-средних.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для сегментации клиентов интернет-магазина. Используется метод k-средних по двум признакам: средний чек и частота покупок за месяц. Число кластеров задано — 3. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения исходных данных, результатов кластеризации и центров кластеров?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			2. Как вы организуете хранение центров кластеров между запусками программы? 3. Почему важно, чтобы структура данных позволяла быстро находить всех клиентов одного кластера? Развёрнутый ответ: 1. Таблица «Клиенты»: номер клиента, средний чек, частота покупок. Таблица «Кластеры»: номер кластера, координата_центра_чек, координата_центра_частота. Таблица «Принадлежность»: номер клиента, номер кластера. 2. Центры кластеров будут храниться в таблице «Кластеры» как отдельные записи. При каждом новом запуске программа может либо использовать старые центры (для продолжения работы), либо пересчитать их заново и обновить таблицу. 3. Чтобы маркетолог мог быстро выбрать всех клиентов из нужной группы (например, «VIP-клиенты») и отправить им персональное предложение. Для этого в таблице «Принадлежность» должен быть индекс по полю «номер кластера», что обеспечит быстрый поиск — это ключевой аспект проектирования эффективной структуры данных. **Вариант 2.** Компания анализирует показатели работы сотрудников отдела продаж. Признаки: число звонков в день и сумма заключённых сделок. Методом k-средних выделены три группы: «активные лидеры», «стабильные исполнители», «требующие поддержки». Задание:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			1. Как вы спроектируете структуру базы данных, чтобы результаты кластеризации можно было использовать в ежедневной работе руководителя? 2. Какие ограничения или правила вы зададите в таблицах, чтобы избежать ошибок при обновлении данных? 3. Почему важно хранить не только номер кластера, но и значения центров кластеров? Развёрнутый ответ: 1. Таблица «Сотрудники»: табельный номер, ФИО, число звонков, сумма сделок. Таблица «Группы продавцов»: номер группы, название группы, центр_звонков, центр_сделок. Таблица «Распределение»: табельный номер, номер группы, дата расчёта. Руководитель сможет видеть состав каждой группы и её «идеальный» профиль (центр). 2. В таблице «Сотрудники» установить: - число звонков ≥ 0, - сумма сделок ≥ 0. В таблице «Распределение» сделать связь по табельному номеру и запретить удалять сотрудника, если он есть в распределении. 3. Центры кластеров показывают типичный профиль для каждой группы. Это помогает: — объяснить, почему сотрудник попал в ту или иную группу, — ставить цели («стремись к центру группы лидеров»), — оценивать изменения со временем. Без хранения центров кластеры превращаются в «чёрные ящики», что снижает ценность анализа

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
Тема 4.9 Метод нечетких к-средних	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Банк хочет сегментировать клиентов по двум признакам: доход и сумма вклада. Однако многие клиенты находятся «на границе» между группами: например, у них высокий доход, но небольшой вклад, или наоборот. Руководство решило использовать метод нечётких k-средних, чтобы учесть эту неопределённость. Задание: 1. Чем отличается результат нечёткой кластеризации от классической в этой ситуации? 2. Как степень принадлежности клиента к разным группам может помочь в принятии решений? 3. Как такая модель может быть использована для автоматизации персональных предложений? Развёрнутый ответ: 1. В классической кластеризации каждый клиент попадает только в одну группу (например, «надёжный» или «рискованный»). В нечёткой — клиент получает долю участия в каждой группе: например, 70% — «надёжный», 30% — «средний». Это точнее отражает реальную картину. 2. Если клиент имеет высокую степень принадлежности к группе «VIP», но и заметную — к группе «неактивный», это сигнал: он потенциально ценен, но требует внимания. Менеджер может предложить ему специальные условия, чтобы удержать. 3. Система автоматически рассчитывает степени принадлежности при поступлении новых данных. Если степень принадлежности к «VIP» превышает 60%, клиенту отправляется предложение о премиум-обслуживании. Если же он «размыт» между группами,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			ему предлагают консультацию. Такой подход повышает точность маркетинга и эффективность взаимодействия. **Вариант 2.** Компания анализирует поведение пользователей мобильного приложения. Признаки: время в приложении и частота покупок. Некоторые пользователи ведут себя неоднозначно: долго сидят, но редко покупают, или наоборот. Классическая кластеризация даёт грубые категории, которые не отражают эту сложность. Задание: 1. Почему нечёткая кластеризация лучше подходит для этой задачи? 2. Как можно использовать значения степеней принадлежности для расчёта эффективности маркетинговых кампаний? 3. Как это повышает надёжность автоматизированной системы анализа? Развёрнутый ответ: 1. Потому что поведение пользователей часто не укладывается в чёткие рамки. Нечёткий метод позволяет каждому пользователю быть «частично активным», «частично пассивным», что даёт более тонкое понимание аудитории. 2. Можно рассчитать среднюю степень принадлежности к группе «платящих» до и после рекламной кампании. Если она выросла — кампания сработала. Это точнее, чем просто считать число новых покупателей, так как учитывает изменение склонности к покупке даже у тех, кто пока не совершил сделку.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			3. Система перестаёт делать жёсткие выводы вроде «этот пользователь неинтересен». Вместо этого она даёт градацию уверенности, что снижает риск ошибок. Например, пользователь с 40% принадлежности к «платящим» всё ещё может стать клиентом — и система не исключит его из рассылки. Это делает автоматизированный анализ гибче и надёжнее
Тема 5.2 Логистическая регрессия.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для банка, которая будет автоматически решать, выдавать ли кредит новому клиенту. Используется логистическая регрессия с признаками: возраст, доход, наличие работы, кредитная история. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения исходных данных, коэффициентов модели и результатов решения? 2. Как вы организуете интерфейс подачи заявки, чтобы данные корректно передавались в модель? 3. Почему важно хранить не только решение («да/нет»), но и вероятность одобрения кредита? Развёрнутый ответ: 1. Таблица «Заявки на кредит»: номер заявки, возраст, доход, наличие работы (да/нет), кредитная история (баллы). Таблица «Модель логистической регрессии»: название признака, коэффициент, дата последнего обучения. Таблица «Решения»: номер заявки, вероятность одобрения, итоговое решение (да/нет), дата расчёта. 2. В программе сделать форму с полями: возраст — целое число от 18 до 80,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			доход — положительное число, работа — переключатель «да/нет», кредитная история — выпадающий список («отличная», «хорошая», «плохая» → преобразуется в баллы). Некорректные данные не принимаются. 3. Вероятность позволяет: — оператору увидеть, насколько решение надёжно (например, 95% — уверенно «да», 52% — пограничный случай), — системе направлять спорные заявки на ручную проверку, — аналитику оценивать качество модели со временем. Без вероятности система становится «чёрным ящиком». **Вариант 2.** Интернет-магазин хочет предсказывать, оформит ли посетитель заказ во время текущего сеанса. Признаки: время на сайте, число просмотренных товаров, источник трафика, наличие товара в корзине. Для этого применяется логистическая регрессия. Задание: 1. Как вы спроектируете структуру базы данных, чтобы поддерживать работу модели в реальном времени? 2. Какие ограничения вы зададите в таблицах, чтобы избежать ошибок (например, отрицательного времени на сайте)? 3. Как результат модели будет использоваться в программном интерфейсе сайта для повышения продаж? Развёрнутый ответ: 1. Таблица «Сессии»: номер сессии, время на сайте (сек), число просмотренных товаров, источник трафика, есть в корзине (да/нет).

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Таблица «Модель прогноза»: признак, коэффициент, порог вероятности для срабатывания акции. Таблица «Прогнозы»: номер сессии, вероятность заказа, рекомендованное действие. 2. В таблице «Сессии» установить правила: время на сайте ≥ 0, число просмотренных товаров ≥ 0, источник трафика — только из справочника (поиск, соцсети, прямой заход). 3. Если вероятность оформления заказа ниже 30%, сайт автоматически показывает всплывающее окно со скидкой 5%. Это действие фиксируется в таблице «Прогнозы». Такой подход требует чёткого взаимодействия между моделью, базой данных и интерфейсом сайта — что и является сутью проектирования программных решений
Тема 5.4 Метрики качества классификации. Регуляризация в задаче регрессии.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом.** **Вариант 1.** Вы разрабатываете программу для прогноза спроса на товар (регрессия) и одновременно систему для выявления мошеннических заказов (классификация). Для регрессии применяется регуляризация, а для классификации — расчёт метрик качества. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения: результатов регрессионной модели с регуляризацией, предсказаний классификатора, метрик его качества? 2. Почему при регуляризации важно хранить исходные и стандартизированные значения признаков отдельно? 3. Как хранение метрик качества (точность, полнота) поможет улучшить работу системы?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			Развёрнутый ответ: 1. Таблица «Регрессия — данные»: дата, признаки (исходные и стандартизированные), фактический спрос, прогноз. Таблица «Регрессия — модель»: название признака, коэффициент, тип регуляризации (гребневая/Lasso), сила регуляризации. Таблица «Классификация — прогнозы»: номер заказа, вероятность мошенничества, решение (да/нет). Таблица «Метрики качества»: дата расчёта, точность, полнота, F1-мера. 2. Потому что регуляризация требует, чтобы все признаки были приведены к одному масштабу (стандартизированы). Но для интерпретации и отчётов нужны исходные значения. Хранение обоих вариантов позволяет и корректно обучать модель, и понятно показывать результаты пользователю. 3. Если, например, полнота низкая, это значит, что система пропускает много реальных случаев мошенничества. Это сигнал о необходимости перенастройки модели. Регулярное хранение метрик позволяет отслеживать качество во времени и принимать решения об обновлении моделей — что невозможно без продуманной структуры данных **Вариант 2.** Компания использует две модели: — регрессионную (с регуляризацией) для прогноза дохода, — классификационную для определения, будет ли клиент отменять заказы. Обе модели работают в единой аналитической системе. Задание:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Задания к практическому занятию
			1. Как вы спроектируете структуру базы данных, чтобы обе модели могли использовать одни и те же исходные данные, но хранить свои результаты отдельно? 2. Какие поля обязательны в таблице «Метрики качества», чтобы можно было сравнивать разные версии классификатора? 3. Почему при использовании регуляризации важно фиксировать параметр «сила регуляризации» в структуре данных? Развёрнутый ответ: 1. Создадим: общую таблицу «Исходные данные»: номер клиента, доход, число заказов, частота отмен и т.д., таблицу «Регрессия — результаты»: номер клиента, прогноз дохода, остаток, версия модели, таблицу «Классификация — результаты»: номер клиента, прогноз отмены, вероятность, версия модели. Общие данные не дублируются, а результаты хранятся раздельно. 2. В таблице «Метрики качества» обязательно указать: дата расчёта, версия модели, набор данных (обучающий/тестовый), точность, полнота, F1-мера. Без этого сравнение будет некорректным. 3. Потому что качество регрессионной модели сильно зависит от силы регуляризации. Если не сохранить это значение, нельзя будет воспроизвести модель или понять, почему изменились прогнозы. Это критически важно для надёжности системы и соответствует требованиям к проектированию структур данных.

Критерии оценивания практических занятий:

Оценка	Критерии оценивания
отлично	Выставляется, если обучающийся умеет увязывать теорию с практикой (решает задачи, формулирует выводы, умеет пояснить полученные результаты), владеет понятийным аппаратом, полно и глубоко овладел материалом по заданной теме, обосновывает свои суждения и даёт правильные ответы на вопросы преподавателя
хорошо	Выставляется, если обучающийся умеет увязывать теорию с практикой (решает задачи и формулирует выводы, умеет пояснить полученные результаты), владеет понятийным аппаратом, полно и глубоко овладел материалом по заданной теме, но содержание ответов имеют некоторые неточности и требуют уточнения и комментария со стороны преподавателя
удовлетворительно	Выставляется, если обучающийся знает и понимает материал по заданной теме, но изложение неполное, непоследовательное, допускаются неточности в определении понятий, студент не может обосновать свои ответы на уточняющие вопросы преподавателя
неудовлетворительно	Выставляется, если обучающийся допускает ошибки в определении понятий, искажающие их смысл, беспорядочно и неуверенно излагает материал. Делает ошибки в ответах на уточняющие вопросы преподавателя

4.3. ТИПОВЫЕ ЗАДАНИЯ ДЛЯ КОНТРОЛЬНЫХ РАБОТ

Контрольные работы содержат несколько практических заданий по индивидуальным вариантам, в полном объеме охватывающих изученный материал по указанной теме. Выполнение контрольных работ позволяет определить результат освоения компетенций по дисциплине в рамках рассматриваемой темы, оцениваемый с помощью соответствующих индикаторов достижения компетенций.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
Раздел №2. «Множественный корреляционно-регрессионный анализ. Классическая модель множественной регрессии» Тема 2.1 Корреляционно-регрессионный анализ. Модель множественной регрессии.	ПК-15	ИПК-15.2	**Контрольная работа №1.** **Задание 1. (Темы 2.1, 2.3)** Аналитик строит модель зависимости прибыли компании от трёх факторов: — расходы на рекламу (тыс. руб.), — число клиентов (тыс. чел.), — средний чек (руб.). После расчётов получены: — коэффициент детерминации R^2=0,82 — множественный коэффициент корреляции R=0,91. Вопросы: 1. Что показывают эти значения? Достаточно ли их, чтобы считать модель качественной?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
Тема 2.2 Фиктивные переменные. Критерий Чоу проверки однородности. Тема 2.3 Множественные коэффициенты корреляции. Тема 2.4 Статистический анализ адекватности модели множественной регрессии. Тема 2.5 Мультиколлинеарность факторов.			2. Какие ещё проверки необходимо выполнить, прежде чем использовать модель для прогноза? 3. Как результаты этой модели могут быть использованы руководством для автоматизации планирования маркетингового бюджета? Ответ: 1. R=0,91 означает, что между прибылью и совокупностью трёх факторов существует очень сильная линейная связь. R^2=0,82 говорит о том, что 82% изменчивости прибыли объясняется моделью. Это хороший показатель, но недостаточный для окончательного вывода о качестве модели. 2. Необходимо дополнительно проверить: — статистическую значимость коэффициентов (t-тесты), — общую значимость модели (F-тест), — наличие мультиколлинеарности между факторами, — выполнение предпосылок МНК: гомоскедастичность, нормальность и независимость остатков. Без этих проверок модель может давать смещённые или неэффективные оценки. 3. Руководство может заложить модель в систему автоматического планирования: — при задании целевой прибыли система рассчитает требуемые расходы на рекламу и ожидаемое число клиентов, — можно оптимизировать бюджет: найти такое сочетание факторов, при котором прибыль максимальна при фиксированных затратах. Это позволяет объективно и быстро принимать решения, повышая экономическую эффективность маркетинга. **Задание 2. (Темы 2.2, 2.4)**

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			Компания работает в двух регионах — Северном и Южном. Есть подозрение, что зависимость прибыли от рекламы отличается в этих регионах. Для проверки используется критерий Чоу. Данные: • Общая модель (все данные): $SSR_{общ}=1200$ • Модель по Северу: $SSR_1=500$ • Модель по Югу: $SSR_2=400$ • Число наблюдений: всего 100, в каждом регионе — по 50 • Число параметров модели: 3 (включая константу) Вопросы: 1. Сформулируйте нулевую гипотезу критерия Чоу. 2. Рассчитайте значение F-статистики Чоу. 3. Если гипотеза однородности отвергается, как это повлияет на проектирование системы автоматизированного управления рекламой в разных регионах? (Формула: $F=\frac{SSR_{общ}-(SSR_1+SSR_2))/k}{SSR_1+SSR_2)/(n-2k)}$, где k — число параметров, n — общее число наблюдений) Ответ: 1. Нулевая гипотеза H_0: модели для Северного и Южного регионов одинаковы, то есть выборки однородны, и можно использовать одну общую модель. 2. Подставим значения в формулу: ○ $SSR_{общ}=1200$ ○ $SSR_1+SSR_2=500+400=900$ ○ k=3, n=100 $F=\frac{(1200-900)/k}{900/(100-2*3)}=\frac{300)/3}{900/94}=\frac{(100)}{9,57}\approx 10,45$ Критическое значение F при у значимости 5%, $df_1=3$,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			df_2=94 ≈ 2,7. Поскольку 10,45 > 2,7, гипотеза H_0 отвергается. 3. Это означает, что одна и та же модель не подходит для обоих регионов. Система автоматизированного управления должна: — использовать две разные модели (по одной на регион), — при планировании бюджета учитывать региональные особенности (например, в одном регионе реклама эффективнее, чем в другом). Это повысит точность прогнозов и экономическую эффективность рекламных кампаний. **Задание 3. (Тема 2.5)** При построении модели зависимости стоимости жилья от следующих факторов: — общая площадь, — жилая площадь, — площадь кухни, — этаж, обнаружена сильная корреляция между площадями (коэффициенты > 0,95). Вопросы: 1. Как мультиколлинеарность может исказить интерпретацию коэффициентов регрессии? Приведите пример возможного парадоксального результата. 2. Почему такая модель может дать ненадёжные прогнозы при автоматической оценке стоимости новых объектов? 3. Предложите два способа устранения мультиколлинеарности и назовите, какой из них предпочтительнее в системе онлайн-оценки недвижимости. Ответ:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			1. При сильной мультиколлинеарности коэффициенты становятся нестабильными и плохо интерпретируемыми. Например, коэффициент при «жилой площади» может оказаться отрицательным, хотя логично, что чем больше жилая площадь — тем дороже квартира. Это происходит потому, что модель «путает» вклад сильно связанных переменных. 2. Небольшие изменения во входных данных (например, ошибка в указании площади) могут привести к резким скачкам в прогнозе, так как коэффициенты имеют завышенную дисперсию. Это делает систему ненадёжной для пользователей. 3. Способы устранения: — исключить один или два из сильно коррелирующих признаков (например, оставить только «общую площадь»), — объединить их в один показатель (например, «удельная стоимость за кв. м» или «соотношение жилой площади к общей»). Для онлайн-системы предпочтительнее первый способ: он проще, быстрее и даёт более устойчивые прогнозы, что критично для массового использования.
Раздел №3. «Обобщенная линейная модель множественной регрессии. Регрессионные модели с автокоррелированными остатками» Тема 3.1 Обобщенная линейная модель множественной	ПК-15	ИПК-5.2	**Контрольная работа №2.** **Задание 1 (Темы 3.1, 3.2, 3.3)** Аналитик строит модель зависимости ежемесячных расходов на рекламу от: — числа клиентов, — региона (центр/периферия), — типа канала (онлайн/офлайн). Диагностика остатков выявила гетероскедастичность: у крупных клиентов разброс расходов значительно выше. Вопросы:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
регрессии. Обобщенный МНК Тема 3.2 Гетероскедастичность Методы выявления гетероскедастичности. Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности. Тема 3.4 Регрессионные модели с автокоррелированными остатками. Тема 3.5 Устранение автокорреляции			1. Как гетероскедастичность влияет на надёжность выводов по модели? 2. Какой метод следует использовать для коррекции и как определить веса? 3. Какие таблицы и поля нужно создать в базе данных, чтобы поддерживать работу скорректированной модели? Ответ: 1. Гетероскедастичность приводит к некорректному расчёту стандартных ошибок, из-за чего t-статистики искажаются. Это может привести к ложному выводу о значимости факторов (например, признать «регион» важным, хотя он не влияет). 2. Следует применить взвешенный МНК. Веса можно определить как обратную величину к дисперсии ошибки. Если дисперсия пропорциональна числу клиентов, то: $W_i \frac{1}{\text{число клиентов}_i}$ 1. Это уменьшит влияние «зашумленных» крупных клиентов. 2. В базе данных создадим: таблицу «Реклама»: месяц, число клиентов, регион, канал, расходы; таблицу «Веса»: месяц, вес (рассчитан по формуле); таблицу «Модель_взвешенная»: название признака, коэффициент, дата расчёта. Такая структура позволяет автоматически пересчитывать веса и обновлять модель. **Задание 2 (Темы 3.4, 3.5)** При анализе ежедневных продаж выявлена положительная автокорреляция остатков: если сегодня продажи выше прогноза, то завтра они тоже чаще выше. Вопросы:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			1. Почему автокорреляция делает обычную регрессию непригодной для планирования закупок? 2. Предложите два способа устранения автокорреляции и назовите, какой из них проще реализовать в программной системе. 3. Как вы спроектируете структуру данных, чтобы поддерживать выбранный метод? Ответ: 1. Автокорреляция означает, что модель не учитывает динамику спроса (например, инерцию или тренд). Прогнозы становятся систематически смещёнными, а доверительные интервалы — уже, чем должны быть. Это приведёт к ошибкам в планировании закупок: либо дефицит, либо избыток товара. 2. Способы устранения: — включение лаговой переменной (продажи за предыдущий день); — применение ОМНК с учётом AR(1)-структуры ошибок. Первый способ проще реализовать: он требует только добавления одного признака и не нуждается в сложных расчётах ковариационной матрицы. 3. В таблице «Продажи» предусмотреть поля: ○ дата, ○ объём продаж, ○ продажи_вчерашний_день (заполняется автоматически при загрузке данных). При построении модели система будет использовать оба поля. Это обеспечивает простоту и надёжность.
Раздел №4 «Кластерный анализ»	ПК-15	ИПК-15.2	**Контрольная работа №3.** **Задание 1 (Темы 4.1, 4.2, 4.7, 4.8)**

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
Тема 4.1 Кластерный анализ. Основные этапы кластерного анализа. Тема 4.2 Метрики расстояний. Свойства кластеров. Методы кластерного анализа. Тема 4.3 Расстояния между кластерами. Методы группировки кластеров. Тема 4.4 Иерархические методы кластеризации. Агломеративные методы. Тема 4.5 Иерархические дивизимные методы кластеризации. Тема 4.6 Примеры реализации иерархических методов кластеризации. Тема 4.7 Итеративные методы. Метод к-средних. Тема 4.8 Метод к-средних. Критерии качества кластеризации. Тема 4.9 Метод нечетких к-средних			Компания хочет сегментировать клиентов интернет-магазина по двум признакам: средний чек и частота покупок. Рассматриваются два подхода: — метод k-средних, — иерархическая кластеризация (агломеративная). Вопросы: 1. Какие этапы первичной обработки данных обязательны перед применением этих методов? 2. Почему важно стандартизировать признаки в этой задаче? 3. Какой критерий качества вы используете для выбора числа кластеров в методе k-средних и почему? Ответ: 1. Обязательные этапы: — удаление или коррекция явных ошибок (например, отрицательный чек), — обработка пропусков (например, замена медианой), — приведение категориальных признаков к числовому виду (если есть), — стандартизация количественных признаков. 2. Средний чек измеряется в рублях (тысячи), а частота — в разах за месяц (единицы). Без стандартизации признак с большим масштабом (чек) будет доминировать, и кластеры будут формироваться только по нему, игнорируя частоту. Стандартизация делает признаки равноправными. 3. Использую метод «локтя» на основе суммы квадратов внутри кластеров (WCSS). Он позволяет визуально определить точку, после которой уменьшение WCSS замедляется — это и есть оптимальное число кластеров. Также можно использовать коэффициент силуэта, который учитывает и компактность, и раздельность кластеров.

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			Задание 2 (Темы 4.3, 4.4, 4.5, 4.6) Аналитик строит иерархическую кластеризацию для группировки товаров по признакам: цена, популярность, сезонность. Возникает выбор между агломеративным и дивизимным подходом. Вопросы: 1. В чём принципиальное различие между агломеративным и дивизимным методами? 2. Как выбор метода расстояния между кластерами (например, «ближайший сосед» vs «дальний сосед») влияет на форму получаемых групп? 3. В какой ситуации дивизимный метод предпочтительнее, и как это связано с бизнес-логикой? Ответ: 1. Агломеративный метод начинается с того, что каждый объект — отдельный кластер, и далее объединяет самые близкие группы. Дивизимный метод начинается с одного кластера, включающего все объекты, и последовательно делит его на подгруппы. 2. Метод «ближайшего соседа» (одиночная связь) склонен образовывать вытянутые, цепочкообразные кластеры, так как объединяет даже через отдельные близкие точки. Метод «дальнего соседа» (полная связь) формирует компактные, шарообразные кластеры, так как требует близости всех объектов. 3. Дивизимный метод предпочтителен, когда бизнес-логика предполагает иерархию решений «от общего к частному». Например, сначала разделить товары на «сезонные» и «несезонные», затем каждую группу — по

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			популярности. Это соответствует мышлению менеджера и делает результаты более интерпретируемыми. **Задание 3 (Темы 4.7, 4.9)** HR-отдел хочет сегментировать сотрудников по двум признакам: стаж и оценка эффективности. Некоторые сотрудники находятся «на границе» между группами (например, высокий стаж, но средняя эффективность). Вопросы: 1. Чем метод нечётких k-средних отличается от классического метода k-средних? 2. Как степень принадлежности к разным кластерам может помочь в управлении персоналом? 3. Почему в этой ситуации нечёткий метод даёт более полезные результаты? Ответ: 1. В классическом методе каждый сотрудник относится только к одному кластеру. В нечётком методе сотрудник имеет степень принадлежности ко всем кластерам (например, 60% — «опытные лидеры», 40% — «требующие развития»). 2. Степени принадлежности позволяют: — выявлять пограничные случаи, требующие индивидуального подхода, — оценивать потенциал роста (например, сотрудник с 30% принадлежности к «лидерам» может стать им при поддержке), — формировать гибкие программы развития, а не жёсткие категории. 3. Потому что поведение и характеристики сотрудников часто не укладываются в чёткие рамки. Нечёткий метод

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			отражает эту реальную неопределённость, тогда как классический метод искусственно «принуждает» к однозначной классификации, теряя важную информацию. Это делает HR-решения точнее и гуманнее.
Раздел №5 «Методы классификации» Тема 5.1 Понятие методов классификации. Тема 5.2 Логистическая регрессия. Тема 5.3 Метрики качества классификации. Тема 5.4 Метрики качества классификации. Регуляризация в задаче регрессии. Тема 5.5 Методы деревьев решений. Тема 5.6 Ансамбли деревьев решений: Случайный лес. Тема 5.7 Ансамбли деревьев решений: Градиентный бустинг.	ПК-15	ИПК-15.2:	**Контрольная работа №4.** **Задание 1 (Темы 5.1, 5.2, 5.3)** Банк разрабатывает систему для автоматического решения о выдаче кредита. Используется логистическая регрессия с признаками: доход, возраст, кредитная история. Результат — вероятность дефолта. Вопросы: 1. Почему логистическая регрессия подходит для этой задачи, а не, например, линейная регрессия? 2. Какие метрики качества вы будете использовать для оценки модели и почему? 3. Какие таблицы и поля вы создадите в базе данных, чтобы хранить данные, модель и результаты? Ответ: 1. Логистическая регрессия предсказывает вероятность принадлежности к классу («дефолт»/«нет»), а не числовое значение. Её результат всегда лежит в диапазоне от 0 до 1, что соответствует смыслу вероятности. Линейная регрессия может дать значения вне этого диапазона, что бессмысленно в задаче классификации. 2. Основные метрики: — Точность (Precision) — доля реальных дефолтов среди всех, кого модель пометила как рискованных (важно, чтобы не отказать надёжным клиентам); — Полнота (Recall) — доля выявленных дефолтов среди всех реальных дефолтов (важно не пропустить рискованных заёмщиков); — F1-мера — среднее гармоническое точности и полноты,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			особенно полезна при дисбалансе классов. Также можно использовать ROC-AUC для оценки качества модели в целом. 3. В базе данных создадим: ○ таблицу «Заявки»: номер, доход, возраст, кредитная_история, решение_эксперта; ○ таблицу «Модель_ЛР»: название_признака, коэффициент, дата_обучения; ○ таблицу «Прогнозы»: номер_заявки, вероятность_дефолта, итоговое_решение, дата_расчёта. Такая структура позволяет отслеживать историю решений и обновлять модель. **Задание 2 (Темы 5.5, 5.6, 5.7)** Компания анализирует поведение пользователей мобильного приложения, чтобы предсказать, оформит ли пользователь подписку. Рассматриваются три метода: — дерево решений, — случайный лес, — градиентный бустинг. Вопросы: 1. Чем дерево решений отличается от ансамблевых методов по интерпретируемости и точности? 2. Почему градиентный бустинг часто даёт более точные прогнозы, чем случайный лес? 3. Как вы спроектируете структуру базы данных, чтобы поддерживать все три модели и сравнивать их качество? Ответ: 1. Дерево решений — очень интерпретируемо: его можно представить в виде правил «если–то». Но оно склонно к переобучению и менее точно. Ансамбли (случайный лес, бустинг) объединяют множество деревьев, что повышает

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			точность, но делает модель менее прозрачной («чёрный ящик»). 2. Градиентный бустинг строит деревья последовательно, каждое следующее исправляет ошибки предыдущих. Это позволяет постепенно уменьшать функцию потерь и достигать высокой точности. Случайный лес строит деревья параллельно и независимо, что снижает дисперсию, но не всегда минимизирует ошибку так эффективно. 3. Создадим следующие таблицы: ○ «Пользователи»: id, время_в_приложении, число_просмотров, подписка (да/нет); ○ «Модели»: id_модели, тип (дерево/лес/бустинг), параметры, дата_обучения; ○ «Прогнозы»: id_пользователя, id_модели, вероятность_подписки, прогноз; ○ «Метрики_качества»: id_модели, точность, полнота, F1, AUC. Такая структура позволяет легко сравнивать модели и выбирать лучшую. **Задание 3 (Темы 5.4, 5.2, 5.5)** При построении модели классификации выяснилось, что некоторые признаки сильно коррелируют (например, «доход» и «уровень образования»). Также есть риск переобучения из-за большого числа признаков. Вопросы: 1. Как мультиколлинеарность влияет на логистическую регрессию и деревья решений? 2. Как регуляризация помогает бороться с переобучением в логистической регрессии?

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовой вариант контрольной работы
			3. Какие ограничения или правила вы зададите в структуре данных, чтобы обеспечить устойчивость моделей? Ответ: 1. В логистической регрессии мультиколлинеарность приводит к нестабильности коэффициентов и завышенным стандартным ошибкам, что затрудняет интерпретацию. В деревьях решений мультиколлинеарность не является проблемой, так как алгоритм выбирает один из сильно связанных признаков, игнорируя остальные. 2. Регуляризация (L1 — Lasso, L2 — гребневая) добавляет штраф за большие коэффициенты в функцию потерь. Это уменьшает сложность модели, снижает риск переобучения и может даже автоматически исключать незначимые признаки (в случае L1). 3. В структуре данных зададим: ○ ограничения на допустимые диапазоны (например, доход ≥ 0), ○ справочники для категориальных признаков (чтобы избежать опечаток), ○ обязательное поле «дата обновления» для каждой модели, ○ хранение только одного из сильно коррелирующих признаков (например, оставить «доход», исключить «уровень образования»). Это повысит надёжность и воспроизводимость моделей

Критерии оценивания контрольной работы:

Оценка	Критерии оценивания
отлично	Выставляется, если обучающийся умеет увязывать теорию с практикой (решает задачи, формулирует выводы, умеет пояснить полученные результаты), владеет понятийным аппаратом, полно и глубоко овладел материалом по заданной теме, обосновывает свои суждения и даёт правильные ответы на вопросы преподавателя

хорошо	Выставляется, если обучающийся умеет увязывать теорию с практикой (решает задачи и формулирует выводы, умеет пояснить полученные результаты), владеет понятийным аппаратом, полно и глубоко овладел материалом по заданной теме, но содержание ответов имеют некоторые неточности и требуют уточнения и комментария со стороны преподавателя.
удовлетворительно	Выставляется, если обучающийся знает и понимает материал по заданной теме, но изложение неполное, непоследовательное, допускаются неточности в определении понятий, студент не может обосновать свои ответы на уточняющие вопросы преподавателя
неудовлетворительно	Выставляется, если обучающийся допускает ошибки в определении понятий, искажающие их смысл, беспорядочно и неуверенно излагает материал. Делает ошибки в ответах на уточняющие вопросы преподавателя

4.4. ТИПОВЫЕ ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Самостоятельная работа включает в себя проработку теоретического материала, изучение рекомендуемой литературы, выполнение практико-ориентированных заданий (заполнение таблиц, проведение сравнительного анализа, составление схем и др.), решение практических задач, создание презентаций, написание рефератов, подборка нормативного и иного материала и выполнение других заданий.

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Типовое задание для самостоятельной работы</th></tr>
<tr><td>Раздел №1. «Введение в статистические методы анализа данных».
Тема 1.1 Введение в статистические методы анализа данных.
Тема 1.2 Методы первичной обработки и представления данных.
Тема 1.3 Методы выявления и устранения аномальных наблюдений.</td><td>ПК-15</td><td>ПК-15.1
ИПК-15.2</td><td>Задание 1.:
Заполните таблицу, указав цель и пример действия для каждого этапа первичного анализа данных.
<table>
<tr><td>Этап анализа данных</td><td>Цель этапа</td><td>Пример действия</td></tr>
<tr><td>Сбор данных</td><td></td><td></td></tr>
<tr><td>Очистка данных</td><td></td><td></td></tr>
<tr><td>Визуализация данных</td><td></td><td></td></tr>
<tr><td>Выявление аномалий</td><td></td><td></td></tr>
</table>
Задание 2.:
Сравните два простых метода выявления аномальных наблюдений:
— метод межквартильного размаха (IQR),
— метод трёх сигм (правило 3σ).</td></tr>
</table>

<table>
<tr><th>Тема</th><th>Код и формулировка компетенции</th><th>Индикаторы достижения компетенции</th><th>Типовое задание для самостоятельной работы</th></tr>
<tr><td></td><td></td><td></td><td>Заполните таблицу:
<table>
<tr><td>Критерий сравнения</td><td>Метод IQR</td><td>Метод 3σ</td></tr>
<tr><td>Для каких данных подходит (нормальные/любые)</td><td></td><td></td></tr>
<tr><td>Требуется ли знание среднего и стандартного отклонения?</td><td></td><td></td></tr>
<tr><td>Чувствителен ли к выбросам при расчёте?</td><td></td><td></td></tr>
</table>
Задание 1.:
Вот фрагмент данных из системы учёта ежедневных продаж в магазине:
<table>
<tr><td>Дата</td><td>Сумма продаж (руб.)</td><td>Число чеков</td><td>Регион</td></tr>
<tr><td>10.01.2024</td><td>125 000</td><td>85</td><td>Центр</td></tr>
<tr><td>11.01.2024</td><td>-50 000</td><td>42</td><td>Поволжье</td></tr>
<tr><td>12.01.2024</td><td></td><td>90</td><td>Север</td></tr>
<tr><td>13.01.2024</td><td>999 999</td><td>-5</td><td>Юг</td></tr>
</table>
Выполните первичную обработку данных:
1. Найдите и опишите ошибки или несоответствия в таблице.</td></tr>
</table>

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовое задание для самостоятельной работы
			2. Предложите, как их исправить или обработать (по каждому случаю). 3. Приведите данные к единообразному и корректному виду, готовому для анализа.
Раздел №2. «Множественный корреляционно-регрессионный анализ. Классическая модель множественной регрессии» Тема 2.1 Корреляционно-регрессионный анализ. Модель множественной регрессии. Тема 2.2 Фиктивные переменные. Критерий Чоу проверки однородности. Тема 2.3 Множественные коэффициенты корреляции. Тема 2.4 Статистический анализ адекватности модели множественной регрессии. Тема 2.5 Мультиколлинеарность факторов.	ПК-5.	ИПК-15.1 ИПК-15.2	**Задание 1.:** Перечислите три основные предпосылки классической модели множественной регрессии (МНК). Почему нарушение хотя бы одной из них может привести к некорректной работе программного модуля прогнозирования (например, завышению точности или ошибкам в принятии решений)? **Задание 2.:** Вы разрабатываете программу для анализа продаж по регионам: «Центр», «Север», «Юг». Опишите, как вы: — закодируете регион с помощью фиктивных переменных, — организуете хранение этих данных в таблице базы данных, — обеспечите передачу данных в расчётный модуль для построения регрессионной модели. **Задание 3.:** Спроектируйте структуру таблицы для хранения данных о ежемесячных продажах, включая: — сумму продаж, — расходы на рекламу, — регион («Центр», «Север», «Юг»). Укажите: — тип каждого поля (числовой, текстовый и т.д.),

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовое задание для самостоятельной работы
			— как именно будет представлен регион (в виде текста или фиктивных переменных), — какие ограничения (например, проверка на положительность) вы зададите для обеспечения корректности данных.
Раздел №3. «Обобщенная линейная модель множественной регрессии. Регрессионные модели с автокоррелированными остатками» Тема 3.1 Обобщенная линейная модель множественной регрессии. Обобщенный МНК Тема 3.2 Гетероскедастичность Методы выявления гетероскедастичности. Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности. Тема 3.4 Регрессионные модели с автокоррелированными остатками. Тема 3.5 Устранение автокорреляции	ПК-5.	ИПК-15.1 ИПК-15.2	**Задание 1.:** Назовите, в чём различие между обычным МНК и обобщённым МНК (ОМНК). В каких двух случаях (нарушения предпосылок) обычный МНК даёт ненадёжные результаты, и почему в этих ситуациях следует использовать ОМНК или его частные случаи (например, взвешенный МНК)? **Задание 2.:** При анализе ежедневных продаж выявлена автокорреляция остатков: высокие продажи сегодня часто сопровождаются высокими завтра. Опишите, как вы: — скорректируете модель (укажите конкретный метод), — измените структуру хранения данных (какие поля добавите в таблицу «Продажи»), — организуете автоматический пересчёт прогноза в программной системе. **Задание 3.:** Спроектируйте структуру таблицы для хранения данных о расходах на обслуживание оборудования, если известно, что у дорогого оборудования разброс расходов гораздо выше (гетероскедастичность). Укажите: — основные поля таблицы,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовое задание для самостоятельной работы
			— какое дополнительное поле вы добавите для поддержки взвешенного МНК, — какие ограничения (например, проверка на положительность) вы зададите для обеспечения корректности данных.
Раздел №4 «Кластерный анализ» Тема 4.1 Кластерный анализ. Основные этапы кластерного анализа. Тема 4.2 Метрики расстояний. Свойства кластеров. Методы кластерного анализа. Тема 4.3 Расстояния между кластерами. Методы группировки кластеров. Тема 4.4 Иерархические методы кластеризации. Агломеративные методы. Тема 4.5 Иерархические дивизимные методы кластеризации. Тема 4.6 Примеры реализации иерархических методов кластеризации. Тема 4.7 Итеративные методы. Метод к-средних.	ПК-5.	ИПК-15.1 ИПК-15.2	**Задание 1.:** Перечислите три основных отличия между иерархическими (агломеративными) и итеративными (например, k-средних) методами кластеризации. Назовите, почему при проектировании программного модуля кластерного анализа важно учитывать эти различия (например, при выборе структуры данных или алгоритма визуализации). **Задание 2.:** Вы разрабатываете систему сегментации клиентов с использованием метода k-средних. Опишите, как вы: — организуете хранение исходных данных о клиентах (например, средний чек, частота покупок), — спроектируете таблицу для хранения центров кластеров и принадлежности клиентов к кластерам, — обеспечите возможность обновления модели при поступлении новых данных. **Задание 3.:** Спроектируйте структуру таблиц для системы, реализующей иерархическую агломеративную кластеризацию товаров. Укажите: — какую таблицу вы создадите для хранения исходных признаков товаров,

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовое задание для самостоятельной работы
Тема 4.8 Метод к-средних. Критерии качества кластеризации. Тема 4.9 Метод нечетких к-средних			— как будет представлена дендрограмма (иерархия объединений) в виде таблицы (укажите поля: номер объединения, какие кластеры/объекты объединяются, расстояние), — какие связи (внешние ключи) вы зададите между таблицами.
Раздел №5 «Методы классификации» Тема 5.1 Понятие методов классификации. Тема 5.2 Логистическая регрессия. Тема 5.3 Метрики качества классификации. Тема 5.4 Метрики качества классификации. Регуляризация в задаче регрессии. Тема 5.5 Методы деревьев решений. Тема 5.6 Ансамбли деревьев решений: Случайный лес. Тема 5.7 Ансамбли деревьев решений: Градиентный бустинг.	ПК-15	ПК-15.1 ИПК-15.2	**Задание 1.:** Назовите, в чём принципиальное различие между задачами классификации и регрессии. Приведите по одному примеру каждой задачи из области управления или экономики. Почему логистическая регрессия относится к методам классификации, несмотря на слово «регрессия» в названии? **Задание 2.:** Банк хочет автоматизировать решение о выдаче кредита. Рассматриваются три метода: — логистическая регрессия, — дерево решений, — случайный лес. Для каждого метода укажите: — преимущество с точки зрения автоматизации (например, скорость, интерпретируемость, точность), — тип метрики качества, который вы будете использовать для оценки (например, точность, полнота, AUC), — как результат будет использоваться в системе принятия решений (например, вероятность > 0,7 → одобрить). **Задание 3.:** После внедрения модели классификации мошеннических операций в банке получены следующие данные за месяц:

Тема	Код и формулировка компетенции	Индикаторы достижения компетенции	Типовое задание для самостоятельной работы
			— всего операций: 10 000, — модель пометила как мошеннические: 300, — из них реально мошеннических: 250, — всего реальных мошеннических операций: 280. Рассчитайте: Точность (Precision) — доля реальных мошенников среди тех, кого нашла система. Полноту (Recall) — долю выявленных мошенников от всех реальных. Назовите, почему обе метрики важны для оценки экономической эффективности и надёжности системы. Формулы: Precision = TP / (TP + FP) Recall = TP / (TP + FN) (TP — истинные положительные, FP — ложные положительные, FN — пропущенные)

Критерии оценивания самостоятельной работы:

Оценка	Критерии оценивания
отлично	выставляется если работа носит научно-исследовательский характер, проанализирован и сделан сравнительный анализ нескольких литературных источников, приведены примеры
хорошо	выставляется если проанализирован и сделан сравнительный анализ нескольких литературных источников, приведены примеры
удовлетворительно	выставляется если проведен сравнительный анализ научно-методической литературы, приведены примеры
неудовлетворительно	выставляется если работа прошла проверку на антиплагиат и соответствует требованиям оформления

4.5. ИНДИВИДУАЛЬНЫЕ ЗАДАНИЯ ДЛЯ КУРСОВЫХ РАБОТ

Курсовая работа позволяет выявить степень владения базовыми знаниями, умениями и навыками, необходимыми для обучения, и определить уровень владения новым материалом.

Примерные индивидуальные задания (темы) для курсовых работ:

КУРСОВАЯ РАБОТА НЕ ПРЕДУСМОТРЕНА УЧЕБНЫМ ПЛАНОМ

Критерии оценивания курсовой работы:

Оценка	Критерии оценивания
отлично	Содержание курсовой работы полностью соответствует заданию, содержащемуся в методических указаниях, и плану. Представлены результаты структурированного и логически последовательного обзора литературных и иных источников по теме исследования. Структура курсовой работы логически и методически выдержана. Верно определены исходные данные для расчетов. Все аналитические расчеты выполнены верно, корректно применены методы экономического анализа, не нарушена методика анализа предмета исследования. Все выводы и предложения убедительно аргументированы. При защите курсовой работы обучающийся правильно и уверенно отвечает на вопросы преподавателя, демонстрирует глубокое знание теоретического материала, способен аргументировать собственные утверждения и выводы
хорошо	Содержание курсовой работы полностью соответствует заданию, содержащемуся в методических указаниях, и плану. Представлены результаты структурированного и логически последовательного обзора литературных и иных источников по теме исследования. Структура курсовой работы логически и методически выдержана. Верно определены исходные данные для расчетов. В расчетах допускаются незначительные (не искажающие общего итога оценки) погрешности/ошибки. Большинство выводов и предложений аргументировано, корректно применены методы экономического анализа, не нарушена методика анализа предмета исследования. Оформление курсовой работы и полученные результаты в целом отвечают требованиям, изложенным в методических указаниях. Имеются одна-две несущественные ошибки в использовании терминов, в построенных диаграммах и схемах, в оформлении таблиц. Наличествует незначительное количество грамматических и/или стилистических ошибок. При защите курсовой работы обучающийся правильно и уверенно отвечает на большинство вопросов преподавателя, демонстрирует хорошее знание теоретического материала, но не всегда способен аргументировать собственные утверждения и выводы. При наводящих вопросах преподавателя исправляет ошибки в ответе

удовлетворительно	Содержание курсовой работы полностью соответствует заданию, содержащемуся в методических указаниях, и плану. Результаты обзора литературных и иных источников представлены недостаточно полно, недостаточно логично и последовательно. Верно определены исходные данные для расчетов, но имеются грубые ошибки в расчетах. Аргументация выводов и предложений слабая или отсутствует. Экономические выводы носят констатирующий (описательный) характер. Имеются одно-два существенных отклонений от требований в оформлении курсовой работы. Полученные результаты в целом отвечают требованиям, изложенным в методических указаниях. Имеются одна-две существенных ошибки в использовании терминов, в построенных диаграммах и схемах. Много грамматических и/или стилистических ошибок. При защите курсовой работы обучающийся допускает грубые ошибки при ответах на вопросы преподавателя, демонстрирует слабое знание теоретического материала, в большинстве случаев не способен уверенно аргументировать собственные утверждения и выводы
неудовлетворительно	Содержание курсовой работы не соответствует заданию, содержащемуся в методических указаниях, и плану. Неверно определены исходные данные для расчетов, неверно и не корректно применены методы экономического анализа. Экономические выводы содержат неверную экономическую оценку. Имеются более двух существенных отклонений от требований в оформлении курсовой работы. Большое количество существенных ошибок по сути работы, много грамматических и стилистических ошибок и др. Полученные результаты не отвечают требованиям, изложенным в методических указаниях. При защите курсовой работы обучающийся демонстрирует слабое понимание программного материала, студент не может защитить свои решения, допускает грубые фактические ошибки при ответах на поставленные вопросы или вовсе не отвечает на них. Курсовая работа не представлена преподавателю. Обучающийся не явился на защиту курсовой работы

5. МАТЕРИАЛЫ ДЛЯ ОЦЕНКИ РЕЗУЛЬТАТОВ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Промежуточная аттестация проводится в форме зачета/зачета с оценкой/экзамена.

5.1. Вопросы к зачету/зачету с оценкой/экзамену:

1. Основные этапы эмпирического исследования.
2. Дискретные, непрерывные данные. Пространственная выборка, временной ряд, панельные данные.
3. Аномальные наблюдения. Критерии проверки аномальности наблюдений.
4. Общий вид модели множественной регрессии. Линейная модель множественной регрессии. Матричная форма.
5. Основные предпосылки регрессионного анализа. Метод наименьших квадратов. МНК-оценки параметров множественной линейной регрессии. Теорема Гаусса-Маркова.
6. Статистические свойства МНК оценок. Проверка гипотезы о статистической значимости теоретических коэффициентов регрессии. Доверительные интервалы для теоретических коэффициентов.
7. Частный коэффициент корреляции, множественный коэффициент корреляции. Проверка гипотезы о статистической значимости частного коэффициента корреляции.

8. Коэффициент детерминации. Свойства. Дисперсионное отношение Фишера. Проверка гипотезы о статистической значимости уравнения регрессии.
9. Обобщенная линейная модель множественной регрессии. Обобщенный МНК.
10. Гомоскедастичность, гетероскедастичность. Критерии проверки остатков на гетероскедастичность. Методы устранения гетероскедастичности.
11. Автокорреляция остатков. Критерии проверки автокорреляции. Устранение автокорреляции.
12. Понятие кластерного анализа. Основные этапы кластерного анализа.
13. Основные меры сходства объектов. Расстояние между объектами: основные виды, формулы расчета.
14. Свойства кластера: плотность, дисперсия, размеры, форма, отделимость. Расстояние между кластерами: основные виды, формулы расчета.
15. Иерархические агломеративные методы: понятие, алгоритм реализации.
16. Иерархические дивизимные методы: понятие, алгоритм реализации.
17. Метод k-средних: понятие, алгоритм реализации. Функционалы качества разбиения объектов методом k-средних.
18. Метод нечетких k-средних: понятие, алгоритм реализации. Функционалы качества разбиения объектов методом k-средних.
19. Постановка задачи классификации. Основные методы классификации.
20. Общий вид логистической регрессии. Оценка логистической регрессии. Понятие сбалансированной выборки.
21. Метрики качества классификации.
22. Методы деревьев решений.
23. Понятие ансамблей решений. Случайный лес. Градиентный бустинг.

Критерии оценивания зачета с оценкой/экзамена:

Оценка	Критерии оценивания
зачтено	Обучающийся должен: уметь строить ответ в соответствии со структурой излагаемого вопроса; продемонстрировать прочное, достаточно полное усвоение знаний программного материала; продемонстрировать знание основных теоретических понятий; правильно формулировать определения; последовательно, грамотно и логически стройно изложить теоретический материал; продемонстрировать умения самостоятельной работы с литературой; уметь сделать достаточно обоснованные выводы по излагаемому материалу.
не зачтено	Обучающийся демонстрирует: незнание значительной части программного материала; не владение понятийным аппаратом дисциплины; существенные ошибки при изложении учебного материала; неумение строить ответ в соответствии со структурой излагаемого вопроса; неумение делать выводы по излагаемому материалу.

Критерии оценивания зачета с оценкой/экзамена:

Оценка	Критерии оценивания

отлично	Выставляется, если обучающимся правильно и полностью раскрыто содержание материала в пределах программы, чётко и правильно даны определения и раскрыто содержание понятий, точно использованы научные и технические термины, в ответе использованы ранее приобретённые теоретические знания, сделаны необходимые выводы и обобщения
хорошо	Выставляется, если обучающимся раскрыто основное содержание материала в пределах программы, даны определения и раскрыто содержание понятий, в ответе использованы ранее приобретённые теоретические знания, сделаны необходимые выводы и обобщения, но присутствуют незначительные нарушения в последовательности изложения, имеются одна-две неточности в содержании ответа
удовлетворительно	Выставляется, если обучающимся содержание учебного материала изложено фрагментарно, не всегда последовательно, не даны определения, не раскрыто содержание понятий, или они изложены с ошибками, допускаются ошибки и неточности в использовании научной терминологии, отсутствуют выводы и обобщения из предыдущего материала, или возможны ошибки в их изложении
неудовлетворительно	Выставляется, если обучающимся основное содержание учебного материала не раскрыто, не даются ответы на основные вопросы, допускаются грубые ошибки в определении понятий, в использовании терминологии, отсутствуют выводы и обобщения

6. МАТЕРИАЛЫ ДЛЯ ДИАГНОСТИЧЕСКОЙ РАБОТЫ

Задания для диагностической работы должны обеспечивать оценку полностью или частично сформированных компетенций. Каждое задание должно быть привязано к тому или иному индикатору сформированности компетенций.

При формировании заданий для диагностической работы необходимо использовать тестовые задания следующих типов:

Тип задания 1. Задания закрытого типа на установление соответствия.

Тип задания 2. Задания закрытого типа на установление последовательности.

Тип задания 3. Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных

Тип задания 4. Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных

Тип задания 5. Задания открытого типа с развернутым ответом.

Все типы заданий должны быть представлены не менее одного раза.

<table>
<tr><th>№ п/п</th><th>Тема занятия</th><th>Код компетенции</th><th>Индикатор</th><th>Тип задания</th><th colspan="2">Задание</th><th>Уровень освоения компетенции</th></tr>
<tr><th></th><th></th><th></th><th></th><th></th><th>Вариант 1</th><th>Вариант 2</th><th></th></tr>
<tr><td>1</td><td>Раздел 1. Тема 1.1 Введение в статистические методы анализа данных.</td><td>ПК-15</td><td>ПК-15.1</td><td>Задания закрытого типа на установление соответствия.</td><td>1. становите соответствие между статистическим понятием и его определением.<table><tr><th>Понятие</th><th>Определение</th></tr><tr><td>А. Генеральная совокупность</td><td>1. Подмножество объектов, отобранных из генеральной совокупности для проведения анализа.</td></tr><tr><td>Б. Выборка</td><td>2. Множество всех возможных объектов или наблюдений, относительно которых делается вывод в исследовании.</td></tr><tr><td>В. Статистический признак</td><td>3. Величина, значение которой зависит от случая и может быть описана с помощью вероятностного распределения.</td></tr><tr><td>Г. Случайная величина</td><td>4. Характеристика единицы наблюдения,</td></tr></table></td><td>1. становите соответствие между этапом статистического исследования и его содержанием.<table><tr><th>Этап исследования</th><th>Содержание этапа</th></tr><tr><td>А. Сбор данных</td><td>1. Применение методов описательной статистики, проверка гипотез, построение моделей.</td></tr><tr><td>Б. Представление данных</td><td>2. Формулировка выводов на основе полученных статистических результатов в контексте предметной области.</td></tr><tr><td>В. Анализ данных</td><td>3. Организация систематического получения информации об объектах исследования (опросы, эксперименты, наблюдения).</td></tr></table></td><td>Базовый (1-3 минуты)</td></tr>
</table>

					которая подлежит измерению или регистрации в ходе исследования. Правильный ответ: А–2, Б–1, В–4, Г–3	Г. Интерпретация результатов — 4. Визуализация и табулирование данных для выявления основных закономерностей (графики, таблицы, диаграммы). Правильный ответ: А–3, Б–4, В–1, Г–2	
2	Раздел 1. Тема 1.2 Методы первичной обработки и представления данных.	ПК-15	ИПК-15.1	**Задания закрытого типа на установление последовательности.**	2. Установите правильную последовательность этапов первичной обработки статистических данных. 1. Расчет сводных статистических показателей (среднее, медиана, мода). 2. Сбор «сырых» данных в результате наблюдения или эксперимента. 3. Проверка и очистка данных от ошибок и пропусков. 4. Группировка и упорядочивание данных (например, построение ранжированного ряда). Правильный ответ: 2 → 3 → 4 → 1	2. Установите правильную последовательность действий при подготовке данных для их последующего представления и анализа. 1. Выбор подходящего метода визуализации (гистограмма, диаграмма, график). 2. Кодирование качественных признаков для машинной обработки. 3. Формирование окончательной таблицы данных (датафрейма) со всеми переменными. 4. Проверка полноты и достоверности собранной информации. Правильный ответ: 4 → 2 → 3 → 1	Базовый (1-3 минуты)
3	Раздел 1. Тема 1.3 Методы выявления и устранения аномаль	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа**	При проектировании базы данных для хранения результатов социологического опроса вы обнаружили, что в поле «Возраст респондента» встречаются значения: перечисленных методов наиболее целесообразно использовать на этапе проектирования для предотвращения подобных аномалий в будущем?	При разработке программного модуля для автоматической загрузки данных в аналитическую систему возникла необходимость фильтрации аномальных значений в количественных переменных. Какой из перечисленных подходов наиболее эффективно сочетает простоту реализации и адаптивность к различным распределениям данных?	Базовый (1-3 минуты)

	ных наблюдений.			**из предложенных**	A) Удаление всех строк с аномальными значениями после загрузки данных. B) Применение правила трёх сигм (3σ) ко всем числовым полям. C) Введение ограничений (constraints) на диапазон допустимых значений при создании таблицы (например, CHECK D) Замена аномальных значений средним по выборке. Правильный ответ: C Обоснование: При проектировании структуры базы данных важно заложить механизмы предотвращения некорректных данных на у СУБД, а не только их последующей очистки. Введение ограничений (CHECK-ограничений) — это стандартное средство проектирования реляционных баз данных, которое обеспечивает целостность данных на этапе ввода. Остальные методы (A, B, D) — это постфактумные процедуры очистки, которые не решают проблему на у проектирования структуры.	A) Удаление всех значений, выходящих за пределы интервала [Q1 – 1.5·IQR; Q3 + 1.5·IQR]. B) Удаление всех значений, превышающих среднее более чем на 2 стандартных отклонения. C) Ручная проверка каждого наблюдения экспертами. D) Замена всех экстремальных значений на медиану. Правильный ответ: A Обоснование: Метод межквартильного размаха (IQR) не предполагает нормальности распределения и устойчив к наличию самих выбросов, в отличие от методов, основанных на среднем и стандартном отклонении (вариант B). Это делает его универсальным и надёжным инструментом для первичной автоматизированной очистки данных в программных системах. Его легко реализовать в коде, и он хорошо интегрируется в конвейеры обработки данных, что важно при проектировании структур и логики работы с данными. Варианты C и D либо не масштабируются, либо искажают данные без необходимости.	
4	Раздел 2. Тема 2.1 Корреляционно-регресс	ПК-15	ПК-15.1	**Задания комбинированного типа, предполагающие выбор**	4. При построении модели множественной линейной регрессии для анализа факторов, влияющих на производительность труда сотрудников, исследователь получил высокий коэффициент детерминации (R2=0,92), но при этом некоторые	4. Какие из перечисленных условий (предпосылок) должны выполняться для корректного применения классической модели множественной линейной регрессии и валидной интерпретации её результатов?	Повышенный (3-5 минут)

	ионный анализ. Модель множественной регрессии.			**нескольких ответов из предложенных**	регрессионные коэффициенты оказались статистически незначимыми. Какие из перечисленных действий целесообразно предпринять для улучшения модели? Выберите все верные варианты: A) Удалить из модели незначимые предикторы и перестроить регрессию. B) Увеличить объём выборки, чтобы повысить мощность статистических тестов. C) Заменить все незначимые коэффициенты на нули без пересчёта модели. D) Проверить наличие мультиколлинеарности между независимыми переменными. E) Игнорировать статистическую значимость, так как высокий R^2 гарантирует качество модели. Правильные ответы: A, B, D Обоснование: A: Удаление незначимых предикторов (особенно при подозрении на мультиколлинеарность) — стандартный шаг в построении интерпретируемой и устойчивой модели. B: Небольшой объём выборки может привести к низкой мощности тестов, из-за чего даже реальные эффекты не будут признаны значимыми.	Выберите все верные варианты: A) Линейная зависимость между зависимой переменной и каждой из независимых переменных. B) Нормальное распределение всех независимых переменных. C) Гомоскедастичность (постоянство дисперсии) случайных ошибок. D) Отсутствие автокорреляции остатков. E) Равенство нулю среднего значения зависимой переменной. Правильные ответы: A, C, D Обоснование: Классическая линейная регрессия опирается на ряд ключевых предпосылок (Гаусса–Маркова): A: Модель предполагает линейную связь в параметрах. C: Гомоскедастичность необходима для эффективности и несмещённости оценок МНК. D: Отсутствие автокорреляции остатков (особенно важно для временных рядов) обеспечивает корректность стандартных ошибок. Почему не B и E: B: Независимые переменные могут иметь любое распределение; нормальность требуется только для ошибок (для построения доверительных интервалов и тестов). E: Среднее зависимой переменной не обязано быть нулём; это требование	

					Увеличение выборки — корректный путь. D: Высокая корреляция между предикторами (мультиколлинеарность) завышает стандартные ошибки коэффициентов, делая их незначимыми, несмотря на реальное влияние. Это необходимо проверить (например, через VIF). Почему не C и E: C нарушает логику регрессионного анализа: модель должна быть пересчитана после изменения состава переменных. E — ошибка: высокий R^2 может быть следствием переобучения или мультиколлинеарности и не гарантирует ни адекватности, ни прогностической способности модели.	относится к математическому ожиданию случайной ошибки ($E[\varepsilon] = 0$), а не к Y.	
5	Раздел 2. Тема 2.2 Фиктивные переменные. Критерий Чоу проверки однородности.	ПК-15	ПК-15.1	**Задания открытого типа с развернутым ответом.**	5. Экономист строит модель зависимости заработной платы сотрудника от его стажа работы. В компании работают как мужчины, так и женщины. Чтобы учесть возможное различие в оплате труда между полами, он решил использовать фиктивную переменную. Вопрос: 1. Что такое фиктивная переменная и как её можно задать в этой задаче? 2. Как будет выглядеть уравнение регрессии с учётом этой переменной? 3. Почему важно учитывать такие качественные признаки (например,	5. Аналитик подозревает, что зависимость между расходами на обучение персонала и производительностью труда до пандемии и после пандемии была разной. Он хочет проверить, изменилась ли структура этой зависимости. Вопрос: 1. Как называется статистический тест, который позволяет сравнить две регрессионные модели до и после определённого события? 2. Что означает, если этот тест показал статистически значимый результат? 3. Почему важно проверять такие гипотезы при анализе экономических процессов?	Высокий (5-10 минут)

					пол) при построении экономических моделей? Развернутый ответ: 1. Фиктивная переменная — это числовая переменная (обычно 0 или 1), которая кодирует качественный признак. Например: o D=1, если сотрудник — мужчина, o D=0, если сотрудник — женщина. 2. Уравнение регрессии: Зарплата=$\beta_0+\beta_1\cdot$Стаж+$\beta_2\cdot D+\varepsilon$ Здесь β_2 покажет, на сколько в среднем отличается зарплата мужчин от зарплаты женщин при одинаковом стаже. 3. Если не учитывать такие факторы, модель может давать неточные или несправедливые выводы. Учёт качественных признаков делает модель более реалистичной и полезной для анализа и управления — это часть системного подхода к моделированию.	Развернутый ответ: 1. Это критерий Чоу (тест на структурный сдвиг или однородность). 2. Если результат значим, это означает, что зависимость между переменными действительно изменилась — то есть одна общая модель для всего периода не подходит. Нужно строить отдельные модели для «до» и «после». Экономические и организационные процессы часто меняются под влиянием внешних факторов (кризисы, реформы, технологии). Проверка однородности помогает избежать ошибок в прогнозах и рекомендациях, делая модели более адекватными реальности — это ключевой элемент математического моделирования	
6	Раздел 2. Тема 2.4 Статистический анализ адекватности модели	ПК-15	ПК-15.1	**Задания закрытого типа на установление последовательности.**	6. Установите правильную последовательность этапов статистического анализа адекватности модели множественной регрессии. 1. Проверка статистической значимости отдельных коэффициентов регрессии с помощью t-теста. 2. Оценка общей значимости модели с помощью F-теста.	6. Установите правильную последовательность действий при интерпретации результатов множественной регрессии. 1. Проверить, значим ли F-критерий модели в целом. 2. Убедиться, что остатки модели не нарушают предпосылок (например, нет гетероскедастичности).	Базовый (1-3 минуты)

	множественной регрессии.				3. Анализ остатков на гомоскедастичность, нормальность и автокорреляцию. 4. Расчёт коэффициента детерминации R^2 и скорректированного R^2_{adj}. Правильный ответ: 4 → 2 → 1 → 3	3. Проанализировать значения коэффициентов регрессии и их экономический смысл. 4. Оценить, насколько хорошо модель объясняет зависимую переменную (по R_2). Правильный ответ: 4 → 1 → 3 → 2	
7	Раздел 2. Тема 2.5 Мультиколлинеарность факторов.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных**	7. При разработке программного модуля для прогнозирования спроса на товар вы построили регрессионную модель с двумя предикторами: «цена товара» и «расходы на рекламу». Однако при анализе оказалось, что коэффициент корреляции между этими двумя переменными равен 0,92. Какая проблема может возникнуть при использовании такой модели в программном обеспечении? Варианты ответов: A) Модель будет давать неточные прогнозы из-за высокой дисперсии оценок коэффициентов. B) Программа не сможет сохранить данные в базе из-за нарушения нормальных форм. C) Время выполнения запроса к базе данных значительно увеличится. D) Все коэффициенты модели автоматически станут нулевыми. Правильный ответ: A Обоснование: Высокая корреляция между независимыми переменными (мультиколлинеарность) не искажает	7. При проектировании структуры данных для аналитической системы вы планируете включить в регрессионную модель следующие переменные: «доход клиента», «средний чек», «ежемесячные траты» и «остаток на счёте». Что следует сделать **до** загрузки этих данных в модель, чтобы избежать проблем с качеством анализа? Варианты ответов: A) Удалить все текстовые поля из таблицы. B) Проверить парные корреляции между этими переменными и при необходимости исключить или объединить сильно коррелирующие. C) Обязательно нормализовать все переменные. D) Создать отдельную таблицу для каждой переменной. Правильный ответ: B Обоснование: Перечисленные переменные могут быть сильными линейными функциями друг друга (например, «доход» и «траты» часто связаны), что вызывает мультиколлинеарность. На этапе	Базовый (1-3 минуты)

					прогнозы, но сильно увеличивает стандартные ошибки коэффициентов, делая их нестабильными и трудно интерпретируемыми. Это снижает надёжность модели, особенно если она используется для анализа влияния отдельных факторов. При проектировании аналитического ПО важно учитывать такие статистические аспекты на этапе подготовки данных и выбора признаков. Остальные варианты не связаны с мультиколлинеарностью: B — проблема проектирования БД, C — проблема производительности, D — неверно.	проектирования аналитической системы важно предусмотреть предварительный анализ взаимосвязей между признаками. Это часть подготовки данных и проектирования логики обработки. Нормализация (C) не решает проблему мультиколлинеарности, а создание отдельных таблиц (D) — избыточно и не связано с этой задачей.	
8	Раздел 3. Тема 3.1 Обобщенная линейная модель множественной регрессии. Обобщенный МНК	ПК-15	ПК-15.1	**Задания комбинированного типа, предполагающие выбор нескольких ответов из предложенных**	8. Аналитик строит регрессионную модель, но при диагностике остатков обнаруживает, что дисперсия ошибок не постоянна (присутствует гетероскедастичность). Какие из перечисленных утверждений верны в этой ситуации? Выберите все верные варианты: A) Оценки коэффициентов, полученные обычным МНК, остаются несмещёнными, но становятся неэффективными. B) Стандартные ошибки коэффициентов, рассчитанные обычным способом, могут быть некорректными, что делает t- и F-тесты ненадёжными. C) Необходимо немедленно удалить половину наблюдений из выборки.	8. Какие из перечисленных ситуаций являются основаниями для использования обобщённого МНК (ОМНК) вместо обычного МНК? Выберите все верные варианты: A) Наличие автокорреляции остатков (например, в данных временных рядов). B) Наличие гетероскедастичности (неравной дисперсии ошибок). C) Все независимые переменные принимают только целые значения. D) Зависимая переменная измерена в рублях. E) Известна структура ковариационной матрицы случайных ошибок (например, из теоретических соображений или предварительного анализа). Правильные ответы: A, B, E	Повышенный (3-5 минут)

					D) Можно применить обобщённый метод наименьших квадратов (ОМНК) для получения более надёжных оценок. E) Гетероскедастичность не влияет на качество модели, если R^2 высокий. Правильные ответы: A, B, D Обоснование: A — верно: при гетероскедастичности МНК-оценки несмещённы, но не имеют минимальной дисперсии (неэффективны). B — верно: стандартные ошибки завышаются или занижаются, поэтому проверка гипотез становится некорректной. D — верно: ОМНК (или робастные стандартные ошибки) — один из способов коррекции проблемы. Почему не C и E: C — произвольное и необоснованное действие, которое ухудшит качество модели. E — заблуждение: высокий R^2 не гарантирует адекватность модели; нарушение предпосылок МНК подрывает статистическую достоверность выводов.	Обоснование: ОМНК применяется, когда ковариационная матрица ошибок не является скалярной (т.е. нарушены условия гомоскедастичности и/или отсутствия автокорреляции): A и B — классические случаи нарушения предпосылок МНК. E — необходимое условие для практического применения ОМНК: нужно знать (или оценить) структуру матрицы Ω. Почему не C и D: Тип шкалы переменных (целые, денежные единицы) не влияет на выбор между МНК и ОМНК. Решающим фактором является поведение остатков, а не формат данных.	
9	Раздел 3. Тема 3.2 Гетероскедастичность.	ПК-15	ИПК-15.1	**Задания закрытого типа на установление последовательности.**	9. Установите правильную последовательность действий при проверке регрессионной модели на наличие гетероскедастичности. 1. Построить график зависимости остатков от предсказанных значений (ŷ).	9. Установите правильную последовательность этапов обработки данных в программном модуле, если после построения регрессионной модели обнаружена гетероскедастичность. 1. Преобразовать переменные (например, взять логарифм зависимой переменной).	Базовый (1-3 минуты)

	Методы выявления гетероскедастичности.				2. Рассчитать остатки модели (разницу между фактическими и предсказанными значениями). 3. Применить формальный статистический тест (например, тест Бреуша–Пагана или Уайта). 4. Сделать вывод о наличии или отсутствии гетероскедастичности. Правильный ответ: 2 → 1 → 3 → 4	2. Подтвердить наличие гетероскедастичности с помощью графического анализа и статистического теста. 3. Перестроить модель с использованием робастных стандартных ошибок или ОМНК. 4. Проверить новую модель на выполнение предпосылок (включая гомоскедастичность). Правильный ответ: 2 → 1 → 3 → 4	
10	Раздел 3. Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности.	ПК-15	ИПК-15.1	**Задания комбинированного типа, предполагающие выбор одного правильного ответа из предложенных**	10. При разработке модуля для анализа данных о доходах населения вы обнаружили, что дисперсия ошибок в регрессионной модели растёт с увеличением уровня дохода (гетероскедастичность). Какой из перечисленных подходов наиболее целесообразен для коррекции этой проблемы на этапе проектирования программного обеспечения? Варианты ответов: A) Удалить все наблюдения с высоким доходом. B) Использовать взвешенный метод наименьших квадратов (ВМНК), присваивая меньший вес наблюдениям с высоким доходом. C) Заменить все значения дохода их средним арифметическим. D) Игнорировать проблему, так как она не влияет на прогнозы. Правильный ответ: B	10. Аналитик строит модель зависимости расходов домохозяйства от его дохода. Он замечает, что разброс расходов выше у богатых семей. Чтобы учесть это при оценивании модели, он решает применить взвешенный МНК. Какой принцип следует использовать для выбора весов? Варианты ответов: A) Присвоить всем наблюдениям одинаковый вес, равный 1. B) Присвоить вес, обратно пропорциональный доходу домохозяйства. C) Присвоить вес, прямо пропорциональный доходу домохозяйства. D) Выбрать веса случайным образом. Правильный ответ: B Обоснование: Если дисперсия ошибки растёт с доходом, то дисперсия $\sigma\frac{2}{i} \propto$ Доход$_i$. В взвешенном МНК оптимальные веса выбираются как $w_i = \frac{1}{\sigma\frac{2}{i}}$, то есть обратно пропорционально дисперсии. Следовательно, вес должен быть обратно	Базовый (1-3 минуты)

					Обоснование: Гетероскедастичность означает, что наблюдения с высоким доходом имеют большую дисперсию ошибки, то есть они менее точны. ВМНК решает эту проблему, уменьшая вес таких «зашумленных» наблюдений при оценке коэффициентов. Это позволяет получить более надёжные и эффективные оценки. Такой подход можно заложить в логику аналитического модуля ПО, и знанию методов обработки данных. Остальные варианты либо искажают данные (A, C), либо игнорируют статистическую проблему (D).	пропорционален доходу. Это обеспечивает большее влияние на оценку тех наблюдений, которые измерены точнее (с меньшей дисперсией). Понимание этого принципа важно при проектировании алгоритмов обработки данных в аналитических системах. Варианты A, C и D не учитывают структуру гетероскедастичности и приведут к неэффективным оценкам.	
11	Тема 2.5 Мультиколлинеарность факторов.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом**	11. Экономист строит модель зависимости прибыли компании от трёх показателей: — расходы на рекламу, — число клиентов, — доход от продаж. При анализе выяснилось, что доход от продаж и число клиентов сильно связаны между собой: чем больше клиентов, тем выше доход. Это явление называется мультиколлинеарностью. Задание: 1. Почему наличие мультиколлинеарности может исказить результаты модели, даже если она хорошо описывает данные? 2. Как это может повлиять на принятие управленческого решения (например,	11. Аналитик разрабатывает модель для оценки стоимости квартир. В неё включены следующие факторы: — общая площадь, — жилая площадь, — площадь кухни. Эти три показателя очень сильно коррелируют между собой: обычно, чем больше общая площадь, тем больше и жилая, и кухня. Задание: 1. Как мультиколлинеарность повлияет на интерпретацию коэффициентов модели? 2. Почему такая модель может быть опасной при автоматизации оценки стоимости (например, в онлайн-калькуляторе)? 3. Какие шаги вы предпримете как разработчик, чтобы обеспечить	Высокий (5-10 минут)

					стоит ли увеличивать расходы на рекламу)? 3. Что можно сделать, чтобы устранить или уменьшить влияние мультиколлинеарности? Развёрнутый ответ: 1. При мультиколлинеарности коэффициенты модели становятся нестабильными: их значения могут сильно меняться при небольшом изменении данных. Также стандартные ошибки коэффициентов завышаются, из-за чего факторы могут казаться незначимыми, хотя на самом деле влияют на результат. 2. Если модель покажет, что «доход от продаж» не значим (из-за коллинеарности с числом клиентов), руководство может ошибочно решить, что рост продаж не влияет на прибыль. Это приведёт к неверным выводам и, возможно, к сокращению активностей, которые на самом деле полезны. 3. Можно: — исключить один из сильно связанных факторов (например, оставить только «число клиентов»), — объединить их в один показатель (например, «средний чек» = доход / число клиентов), — использовать методы, устойчивые к коллинеарности (например, регуляризацию).	надёжность системы автоматической оценки? Развёрнутый ответ: 1. Коэффициент при «жилой площади» может получиться отрицательным, хотя логично, что чем больше жилая площадь — тем дороже квартира. Это происходит потому, что модель «путает» вклад каждого фактора из-за их сильной связи. 2. Автоматическая система может начать давать странные оценки: например, предлагать более высокую цену за квартиру с меньшей жилой площадью. Это подорвёт доверие пользователей и приведёт к финансовым потерям. Как разработчик: — проверить парные корреляции между факторами, — удалю или объединю сильно связанные переменные (например, оставлю только «общую площадь»), — протестировать модель на новых данных, — заложить в систему предупреждение, если входные данные нарушают логические связи (например, жилая площадь > общей). Это обеспечит устойчивость и достоверность автоматизированного решения	

					Это повысит надёжность модели и качество принимаемых решений		
12	Тема 3.1 Обобщенная линейная модель множественной регрессии. Обобщенный МНК	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом**	12. Вы анализируете данные о времени обработки заявок в технической поддержке компании. Зависимая переменная — время (в минутах). Факторы: — сложность заявки (баллы), — загруженность оператора (число активных заявок), — тип оборудования клиента. При проверке остатков выяснилось, что дисперсия ошибок растёт с увеличением времени: для сложных заявок разброс времени гораздо больше, чем для простых. Это нарушение условия равной дисперсии (гомоскедастичности). Задание: 1. Почему в этой ситуации обычный метод наименьших квадратов (МНК) может дать ненадёжные результаты? 2. Какой метод следует использовать вместо обычного МНК и почему? 3. Как это повлияет на расчёт показателей эффективности работы службы поддержки? Развёрнутый ответ: 1. При неравной дисперсии (гетероскедастичности) оценки коэффициентов остаются несмещёнными, но их стандартные ошибки рассчитываются неверно. Из-за этого тесты на значимость	12. Компания прогнозирует расходы на обслуживание оборудования в зависимости от: — возраста оборудования (лет), — условий эксплуатации (шкала от 1 до 5), — количества предыдущих ремонтов. Анализ показал, что ошибки модели коррелируют во времени: если в одном месяце расходы были завышены, то и в следующем они тоже выше ожидаемого. Это указывает на автокорреляцию остатков. Задание: 1. Почему автокорреляция делает обычную регрессионную модель непригодной для надёжного прогноза? 2. Какой подход вы выберете для коррекции модели и почему? 3. Как использование обобщённой модели повлияет на формализацию процесса планирования бюджета на обслуживание? Развёрнутый ответ: 1. Автокорреляция означает, что в модели не учтена какая-то закономерность (например, сезонность или износ). Из-за этого прогнозы становятся систематически смещёнными, а доверительные интервалы — уже, чем должны быть. Это ведёт к ошибкам в планировании бюджета. 2. Воспользоваться обобщённым методом наименьших квадратов (ОМНК) с учётом	Высокий (5-10 минут)

					факторов становятся недостоверными, и можно принять ложные выводы (например, признать важный фактор незначимым). 2. Следует использовать обобщённый метод наименьших квадратов (ОМНК). Он учитывает структуру дисперсии ошибок и даёт более точные и надёжные оценки, особенно когда дисперсия зависит от уровня зависимой переменной. 3. Более точная модель позволит правильно оценить влияние каждого фактора на время обработки. Например, можно выявить, что именно загруженность оператора, а не сложность заявки, — главная причина задержек. Это поможет обоснованно перераспределить нагрузку и точнее планировать ресурсы, что повысит объективность расчёта показателей эффективности	структуры автокорреляции (например, предположу, что ошибки следуют модели авторегрессии первого порядка). Это позволит учесть зависимость между соседними наблюдениями и получить более точные оценки. На основе скорректированной модели можно строить реалистичные прогнозы с корректными интервалами неопределённости. Это позволит финансовой службе формализовать процесс бюджетирования: вместо фиксированной суммы закладывать диапазон, зависящий от состояния парка оборудования. Такой подход повышает надёжность и обоснованность расчётов.	
13	Тема 3.3 Взвешенный МНК. Устранение гетероскедастичности.	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	13. Вы разрабатываете программу для анализа расходов домохозяйств. Зависимая переменная — месячные расходы (в рублях). Факторы: доход, число членов семьи, город проживания. Анализ показал: у богатых семей разброс расходов гораздо больше, чем у бедных. Это означает, что дисперсия ошибок модели зависит от уровня дохода (гетероскедастичность). Задание:	13. Компания анализирует издержки на доставку товаров в разные регионы. Зависимая переменная — стоимость доставки. Факторы: расстояние, вес груза, тип дороги. Выяснилось, что в удалённых регионах разброс цен на доставку очень велик, а в центральных — стабилен. Это классический случай гетероскедастичности. Задание:	Высокий (5-10 минут)

					1. Почему в этой ситуации обычная регрессия даст ненадёжные результаты? 2. Как вы измените структуру расчётов в программе, чтобы учесть эту особенность? 3. Какие таблицы и поля вы создадите в базе данных, чтобы поддерживать работу модели с весами? Развёрнутый ответ: 1. Обычная регрессия считает все наблюдения одинаково точными. Но у богатых семей данные «зашумлены» сильнее, поэтому их влияние на модель должно быть меньше. Без учёта этого модель будет смещённой и неустойчивой, особенно при прогнозировании для малообеспеченных семей. 2. В программе нужно реализовать взвешенный метод наименьших квадратов: — рассчитать вес для каждого домохозяйства как обратную величину к дисперсии (например, вес = 1 / доход), — использовать эти веса при расчёте коэффициентов модели. 3. Таблица «Домохозяйства»: номер, доход, число членов, город, расходы. Таблица «Веса»: номер домохозяйства, вес (рассчитан по формуле). Таблица «Модель»: название коэффициента, значение, дата расчёта.	1. Как вы организуете ввод и хранение данных в базе, чтобы подготовить их к применению взвешенного МНК? 2. Какой механизм расчёта весов вы заложите в программу? 3. Как пользователь (логист) будет получать прогноз стоимости доставки с учётом этой модели? Развёрнутый ответ: 1. В таблице «Доставки» предусмотреть поля: регион, расстояние, вес, тип дороги, стоимость. Дополнительно создать справочник «Регионы» с признаком «удалённость» (центр/периферия). Это позволит автоматически определять, для каких записей нужен больший вес. 2. Программа будет рассчитывать вес по формуле: — для центральных регионов: вес = 1, — для удалённых: вес = 0,3 (или другое значение, полученное из данных). Веса можно также рассчитывать как обратную величину к среднему разбросу по региону. 3. Логист вводит параметры заказа (расстояние, вес, регион). Программа автоматически определяет вес, применяет модель и выводит прогнозируемую стоимость. Результат отображается в окне расчёта заказа — это требует чёткого проектирования интерфейса и логики взаимодействия между данными и моделью	

					Такая структура позволяет гибко обновлять веса и пересчитывать модель при поступлении новых данных.		
14	Тема 3.5 Устранение автокорреляции	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	14. Вы разрабатываете систему для прогнозирования ежедневных продаж в магазине. В регрессионную модель входят: количество рекламных объявлений, погода и день недели. После построения модели выяснилось, что ошибки модели коррелируют во времени: если сегодня продажи выше прогноза, то завтра они тоже чаще выше. Это признак автокорреляции остатков. Задание: 1. Почему автокорреляция делает обычную регрессию непригодной для надёжного прогноза? 2. Как вы измените структуру хранения данных, чтобы учесть зависимость между соседними днями? 3. Какие дополнительные поля нужно добавить в таблицу с данными, чтобы реализовать метод устранения автокорреляции? Развёрнутый ответ: 1. Автокорреляция означает, что в модели не учтена внутренняя динамика процесса (например, инерция спроса). Из-за этого прогнозы становятся систематически смещёнными, а оценки точности — оптимистичными. Это	14. Компания анализирует ежемесячные расходы на электроэнергию в зависимости от: площади помещения, числа сотрудников и температуры на улице. Анализ остатков показал положительную автокорреляцию: высокие расходы в одном месяце часто сменяются высокими в следующем. Это связано с сезонностью и инерцией потребления. Задание: 1. Какой способ устранения автокорреляции вы выберете и почему? 2. Как вы спроектируете структуру базы данных, чтобы поддерживать этот метод? 3. Почему важно предусмотреть такую структуру заранее, ещё на этапе проектирования системы учёта? Развёрнутый ответ: 1. Выбрать включение лаговой переменной — то есть добавить в модель расходы на электроэнергию за предыдущий месяц. Это просто, интерпретируемо и эффективно устраняет автокорреляцию, вызванную инерцией процесса. 2. В основной таблице «Энергопотребление» нужно предусмотреть: о поле «месяц» (с возможностью сортировки по возрастанию), о поле «расход_электроэнергии»,	Высокий (5-10 минут)

					особенно опасно при планировании закупок. 2. Нужно хранить данные в строгом хронологическом порядке и обеспечить возможность обращения к предыдущим наблюдениям. Для этого в таблице обязательно должен быть порядковый номер дня или дата, по которой можно однозначно определить предыдущий и следующий период. 3. Следует добавить поле «продажи в предыдущий день» как новый фактор. Также можно создать отдельную таблицу «Остатки модели», где хранить ошибки за каждый день — это понадобится, если использовать методы, учитывающие структуру автокорреляции (например, преобразование Кохрана–Оркатта).	о поле «расход_электроэнергии_предыдущий_месяц» (может заполняться автоматически при загрузке данных). Также, можно создать связь «один ко многим» между месяцами для удобного доступа к истории. 3. Если не заложить такую структуру заранее, потом будет сложно или невозможно добавить лаговые переменные без перестройки всей системы. Это приведёт к дополнительным затратам времени и ресурсов. Правильное проектирование структуры данных с учётом временной зависимости — ключ к надёжной и расширяемой системе.	
15	Тема 4.3 Расстояния между кластерами. Методы группировки кластеров.	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом**	15. Компания проводит сегментацию клиентов по двум признакам: средний чек и частота покупок. Для этого используется иерархическая кластеризация. Возникает вопрос: как определять расстояние между уже образованными группами — по ближайшим объектам, по центрам или как-то ещё? Задание: 1. Опишите два разных способа определения расстояния между кластерами (например, метод ближайшего и дальнего соседа). Чем они отличаются?	15. Аналитик строит иерархическую кластеризацию для группировки товаров по признакам: цена, популярность, сезонность. Руководство хочет использовать результаты для расчёта эффективности товарных групп и планирования закупок. Задание: 1. Почему выбор метода расчёта расстояния между кластерами влияет на инженерные расчёты показателей эффективности? 2. Какой метод вы порекомендуете — среднюю связь или центроидный метод — и почему?	Высокий (5-10 минут)

					2. Как выбор метода повлияет на форму и количество полученных групп? 3. Почему важно заранее выбрать метод группировки при формализации процесса сегментации в информационной системе? Развёрнутый ответ: 1. Метод ближайшего соседа: расстояние между кластерами — это расстояние между самыми близкими объектами из разных групп. Метод дальнего соседа: расстояние — между самыми удалёнными объектами. Первый метод даёт вытянутые, цепочкообразные кластеры, второй — компактные и шарообразные. 2. При использовании метода ближайшего соседа может получиться мало крупных кластеров, даже если данные содержат мелкие обособленные группы. Метод дальнего соседа, наоборот, сохраняет компактность, но может дробить естественные скопления. 3. Если метод не зафиксирован заранее, то при каждом запуске система может давать разные результаты, что сделает сегментацию непредсказуемой. Для автоматизации маркетинговых решений (например, рассылки по группам) нужна устойчивая и воспроизводимая логика, поэтому выбор метода — часть формализации процесса	3. Как вы обеспечите, чтобы результаты кластеризации можно было использовать в расчётах автоматически? Развёрнутый ответ: 1. Разные методы дают разные составы групп. Например, одна и та же группа товаров может включать разные позиции в зависимости от метода. Это напрямую влияет на расчёт средней рентабельности, оборачиваемости или спроса по группе — а значит, и на решения о закупках. 2. Я рекомендую метод средней связи, потому что он устойчив к выбросам и даёт сбалансированные кластеры. Центроидный метод может привести к инверсии (нарушению монотонности расстояний), что затруднит интерпретацию дендрограммы и автоматизацию. 3. Чтобы результаты можно было использовать автоматически, нужно: — сохранять номер кластера для каждого товара в справочной таблице, — фиксировать порог объединения (например, «объединять до расстояния 5»), — хранить правила группировки (метод, метрика) в настройках системы. Это позволит регулярно пересчитывать группы и использовать их в других модулях (например, в модуле планирования закупок), что соответствует требованиям к формализации и расчёту показателей результативности	

16	Тема 4.5 Иерархические дивизимные методы кластеризации.	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	16. Вы разрабатываете программу для анализа ассортимента крупного магазина. В начале все товары считаются одной группой. Затем система должна последовательно делить эту группу на подгруппы (например, сначала — «продукты» и «непродукты», потом — «молочка», «мясо» и т.д.). Для этого используется дивизимный метод кластеризации. Задание: 1. Какие таблицы и поля вы создадите в базе данных, чтобы хранить результаты такого иерархического деления? 2. Как вы организуете логику выбора кластера для деления в программе? 3. Как пользователь (менеджер по ассортименту) будет просматривать и использовать полученную иерархию групп? Развёрнутый ответ: 1. Таблица «Товары»: номер товара, название, признаки (цена, вес, категория и т.д.). Таблица «Кластеры»: номер кластера, номер родительского кластера, уровень вложенности, описание. Таблица «Состав кластера»: номер товара, номер кластера. Такая структура позволяет строить дерево групп любой глубины. 2. Программа будет на каждом шаге:	16. Компания хочет сегментировать своих клиентов, начав с одного общего списка и последовательно разделяя его на всё более однородные подгруппы (например, сначала — по активности, затем — по доходу, затем — по интересам). Выбран дивизимный подход, так как он лучше отражает логику бизнес-анализа «от общего к частному». Задание: 1. Почему дивизимный метод удобнее для бизнес-аналитика, чем агломеративный, в этой ситуации? 2. Как вы спроектируете структуру базы данных, чтобы поддерживать процесс последовательного деления? 3. Какие ограничения вы введёте, чтобы избежать слишком мелких или пустых групп? Развёрнутый ответ: 1. Дивизимный метод начинается с единой целостной картины (все клиенты), что соответствует реальному взгляду руководителя. Аналитик сам решает, по какому признаку важнее всего разделить клиентов на первом шаге, что делает процесс более управляемым и интерпретируемым. 2. Создадим таблицы: о «Клиенты»: номер, признаки (доход, активность и т.д.), о «Группы»: номер группы, номер родительской группы, уровень, правило деления (например, «доход > 100 тыс.»),	Высокий (5-10 минут)

					— рассчитывать меру неоднородности для каждого текущего кластера (например, внутрикластерную дисперсию), — выбирать кластер с наибольшей неоднородностью для деления, — применять метод (например, k-средние с k=2) для разделения на две части. 3. Менеджер увидит древовидное меню в интерфейсе программы: корень — «Все товары», далее — подгруппы. Он сможет: — остановить деление на нужном у , — присвоить названия группам, — экспортировать состав каждой группы для анализа продаж. Это требует чёткого проектирования интерфейса и структуры данных.	о «Состав групп»: номер клиента, номер группы. Связь «группа → подгруппы» реализуется через поле «родительская группа». 3. В программе зададим правила: — минимальный размер группы (например, не менее 10 клиентов), — максимальная глубина дерева (например, не более 5 у й), — запрет на деление, если внутри группы все клиенты одинаковы по ключевым признакам. Это предотвратит бессмысленные разбиения и обеспечит устойчивость системы	
17	Тема 4.7 Итеративные методы. Метод к-средних.	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	17. Вы разрабатываете программу для сегментации клиентов интернет-магазина. Используется метод k-средних по двум признакам: средний чек и частота покупок за месяц. Число кластеров задано — 3. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения исходных данных, результатов кластеризации и центров кластеров? 2. Как вы организуете хранение центров кластеров между запусками программы?	17. Компания анализирует показатели работы сотрудников отдела продаж. Признаки: число звонков в день и сумма заключённых сделок. Методом k-средних выделены три группы: «активные лидеры», «стабильные исполнители», «требующие поддержки». Задание: 1. Как вы спроектируете структуру базы данных, чтобы результаты кластеризации можно было использовать в ежедневной работе руководителя? 2. Какие ограничения или правила вы зададите в таблицах, чтобы избежать ошибок при обновлении данных?	Высокий (5-10 минут)

					3. Почему важно, чтобы структура данных позволяла быстро находить всех клиентов одного кластера? Развёрнутый ответ: 1. Таблица «Клиенты»: номер клиента, средний чек, частота покупок. Таблица «Кластеры»: номер кластера, координата_центра_чек, координата_центра_частота. Таблица «Принадлежность»: номер клиента, номер кластера. 2. Центры кластеров будут храниться в таблице «Кластеры» как отдельные записи. При каждом новом запуске программа может либо использовать старые центры (для продолжения работы), либо пересчитать их заново и обновить таблицу. 3. Чтобы маркетолог мог быстро выбрать всех клиентов из нужной группы (например, «VIP-клиенты») и отправить им персональное предложение. Для этого в таблице «Принадлежность» должен быть индекс по полю «номер кластера», что обеспечит быстрый поиск — это ключевой аспект проектирования эффективной структуры данных.	3. Почему важно хранить не только номер кластера, но и значения центров кластеров? Развёрнутый ответ: 1. Таблица «Сотрудники»: табельный номер, ФИО, число звонков, сумма сделок. Таблица «Группы продавцов»: номер группы, название группы, центр_звонков, центр_сделок. Таблица «Распределение»: табельный номер, номер группы, дата расчёта. Руководитель сможет видеть состав каждой группы и её «идеальный» профиль (центр). 2. В таблице «Сотрудники» установить: - число звонков ≥ 0, - сумма сделок ≥ 0. В таблице «Распределение» сделать связь по табельному номеру и запретить удалять сотрудника, если он есть в распределении. 3. Центры кластеров показывают типичный профиль для каждой группы. Это помогает: — объяснить, почему сотрудник попал в ту или иную группу, — ставить цели («стремись к центру группы лидеров»), — оценивать изменения со временем. Без хранения центров кластеры превращаются в «чёрные ящики», что снижает ценность анализа	
18	Тема 4.9 Метод нечетких к-средних	ПК-15	ИПК-15.2	**Задания открытого типа с развернутым ответом**	18. Банк хочет сегментировать клиентов по двум признакам: доход и сумма вклада. Однако многие клиенты находятся «на границе» между группами: например, у них высокий доход, но небольшой вклад, или	18. Компания анализирует поведение пользователей мобильного приложения. Признаки: время в приложении и частота покупок. Некоторые пользователи ведут себя неоднозначно: долго сидят, но редко покупают, или наоборот. Классическая	Высокий (5-10 минут)

					наоборот. Руководство решило использовать метод нечётких k-средних, чтобы учесть эту неопределённость. Задание: 1. Чем отличается результат нечёткой кластеризации от классической в этой ситуации? 2. Как степень принадлежности клиента к разным группам может помочь в принятии решений? 3. Как такая модель может быть использована для автоматизации персональных предложений? Развёрнутый ответ: 1. В классической кластеризации каждый клиент попадает только в одну группу (например, «надёжный» или «рискованный»). В нечёткой — клиент получает долю участия в каждой группе: например, 70% — «надёжный», 30% — «средний». Это точнее отражает реальную картину. 2. Если клиент имеет высокую степень принадлежности к группе «VIP», но и заметную — к группе «неактивный», это сигнал: он потенциально ценен, но требует внимания. Менеджер может предложить ему специальные условия, чтобы удержать. 3. Система автоматически рассчитывает степени принадлежности при поступлении новых данных. Если степень принадлежности к «VIP»	кластеризация даёт грубые категории, которые не отражают эту сложность. Задание: 1. Почему нечёткая кластеризация лучше подходит для этой задачи? 2. Как можно использовать значения степеней принадлежности для расчёта эффективности маркетинговых кампаний? 3. Как это повышает надёжность автоматизированной системы анализа? Развёрнутый ответ: 1. Потому что поведение пользователей часто не укладывается в чёткие рамки. Нечёткий метод позволяет каждому пользователю быть «частично активным», «частично пассивным», что даёт более тонкое понимание аудитории. 2. Можно рассчитать среднюю степень принадлежности к группе «платящих» до и после рекламной кампании. Если она выросла — кампания сработала. Это точнее, чем просто считать число новых покупателей, так как учитывает изменение склонности к покупке даже у тех, кто пока не совершил сделку. 3. Система перестаёт делать жёсткие выводы вроде «этот пользователь неинтересен». Вместо этого она даёт градацию уверенности, что снижает риск ошибок. Например, пользователь с 40% принадлежности к «платящим» всё ещё может стать клиентом — и система не исключит его из рассылки. Это делает	

					превышает 60%, клиенту отправляется предложение о премиум-обслуживании. Если же он «размыт» между группами, ему предлагают консультацию. Такой подход повышает точность маркетинга и эффективность взаимодействия.	автоматизированный анализ гибче и надёжнее	
19	Тема 5.2 Логистическая регрессия.	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	19. Вы разрабатываете программу для банка, которая будет автоматически решать, выдавать ли кредит новому клиенту. Используется логистическая регрессия с признаками: возраст, доход, наличие работы, кредитная история. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения исходных данных, коэффициентов модели и результатов решения? 2. Как вы организуете интерфейс подачи заявки, чтобы данные корректно передавались в модель? 3. Почему важно хранить не только решение («да/нет»), но и вероятность одобрения кредита? Развёрнутый ответ: 1. Таблица «Заявки на кредит»: номер заявки, возраст, доход, наличие работы (да/нет), кредитная история (баллы). Таблица «Модель логистической регрессии»: название признака, коэффициент, дата последнего обучения.	19. Интернет-магазин хочет предсказывать, оформит ли посетитель заказ во время текущего сеанса. Признаки: время на сайте, число просмотренных товаров, источник трафика, наличие товара в корзине. Для этого применяется логистическая регрессия. Задание: 1. Как вы спроектируете структуру базы данных, чтобы поддерживать работу модели в реальном времени? 2. Какие ограничения вы зададите в таблицах, чтобы избежать ошибок (например, отрицательного времени на сайте)? 3. Как результат модели будет использоваться в программном интерфейсе сайта для повышения продаж? Развёрнутый ответ: 1. Таблица «Сессии»: номер сессии, время на сайте (сек), число просмотренных товаров, источник трафика, есть в корзине (да/нет). Таблица «Модель прогноза»: признак, коэффициент, порог вероятности для срабатывания акции.	Высокий (5-10 минут)

					Таблица «Решения»: номер заявки, вероятность одобрения, итоговое решение (да/нет), дата расчёта. 2. В программе сделать форму с полями: возраст — целое число от 18 до 80, доход — положительное число, работа — переключатель «да/нет», кредитная история — выпадающий список («отличная», «хорошая», «плохая» → преобразуется в баллы). Некорректные данные не принимаются. 3. Вероятность позволяет: — оператору увидеть, насколько решение надёжно (например, 95% — уверенно «да», 52% — пограничный случай), — системе направлять спорные заявки на ручную проверку, — аналитику оценивать качество модели со временем. Без вероятности система становится «чёрным ящиком».	Таблица «Прогнозы»: номер сессии, вероятность заказа, рекомендованное действие. 2. В таблице «Сессии» установить правила: время на сайте ≥ 0, число просмотренных товаров ≥ 0, источник трафика — только из справочника (поиск, соцсети, прямой заход). 3. Если вероятность оформления заказа ниже 30%, сайт автоматически показывает всплывающее окно со скидкой 5%. Это действие фиксируется в таблице «Прогнозы». Такой подход требует чёткого взаимодействия между моделью, базой данных и интерфейсом сайта — что и является сутью проектирования программных решений	
20	Тема 5.4 Метрики качества классификации. Регуляризация в задаче	ПК-5.	ИПК-15.2	**Задания открытого типа с развернутым ответом**	20. Вы разрабатываете программу для прогноза спроса на товар (регрессия) и одновременно систему для выявления мошеннических заказов (классификация). Для регрессии применяется регуляризация, а для классификации — расчёт метрик качества. Задание: 1. Какие таблицы и поля вы создадите в базе данных для хранения:	20. Компания использует две модели: — регрессионную (с регуляризацией) для прогноза дохода, — классификационную для определения, будет ли клиент отменять заказы. Обе модели работают в единой аналитической системе. Задание: 1. Как вы спроектируете структуру базы данных, чтобы обе модели могли использовать одни и те же исходные	Высокий (5-10 минут)

	регрессии.				результатов регрессионной модели с регуляризацией, предсказаний классификатора, метрик его качества? 2. Почему при регуляризации важно хранить исходные и стандартизированные значения признаков отдельно? 3. Как хранение метрик качества (точность, полнота) поможет улучшить работу системы? Развёрнутый ответ: 1. Таблица «Регрессия — данные»: дата, признаки (исходные и стандартизированные), фактический спрос, прогноз. Таблица «Регрессия — модель»: название признака, коэффициент, тип регуляризации (гребневая/Lasso), сила регуляризации. Таблица «Классификация — прогнозы»: номер заказа, вероятность мошенничества, решение (да/нет). Таблица «Метрики качества»: дата расчёта, точность, полнота, F1-мера. 2. Потому что регуляризация требует, чтобы все признаки были приведены к одному масштабу (стандартизированы). Но для интерпретации и отчётов нужны исходные значения. Хранение обоих вариантов позволяет и корректно обучать модель, и понятно показывать результаты пользователю.	данные, но хранить свои результаты отдельно? 2. Какие поля обязательны в таблице «Метрики качества», чтобы можно было сравнивать разные версии классификатора? 3. Почему при использовании регуляризации важно фиксировать параметр «сила регуляризации» в структуре данных? Развёрнутый ответ: 1. Создадим: общую таблицу «Исходные данные»: номер клиента, доход, число заказов, частота отмен и т.д., таблицу «Регрессия — результаты»: номер клиента, прогноз дохода, остаток, версия модели, таблицу «Классификация — результаты»: номер клиента, прогноз отмены, вероятность, версия модели. Общие данные не дублируются, а результаты хранятся раздельно. 2. В таблице «Метрики качества» обязательно указать: дата расчёта, версия модели, набор данных (обучающий/тестовый), точность, полнота, F1-мера. Без этого сравнение будет некорректным. 3. Потому что качество регрессионной модели сильно зависит от силы регуляризации. Если не сохранить это значение, нельзя будет воспроизвести модель или понять, почему изменились	

					3. Если, например, полнота низкая, это значит, что система пропускает много реальных случаев мошенничества. Это сигнал о необходимости перенастройки модели. Регулярное хранение метрик позволяет отслеживать качество во времени и принимать решения об обновлении моделей — что невозможно без продуманной структуры данных	прогнозы. Это критически важно для надёжности системы и соответствует требованиям к проектированию структур данных.	

Критерии оценивания диагностической работы:

Оценка	Критерии оценивания
отлично	от 90 до 100 % правильно выполненных заданий
хорошо	от 70 до 89 % правильно выполненных заданий
удовлетворительно	от 50 до 69 % правильно выполненных заданий
неудовлетворительно	менее 50 % правильно выполненных заданий