

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ЛУГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ ВЛАДИМИРА ДАЛЯ»

Экономический институт
Кафедра экономической кибернетики и прикладной статистики
(наименование кафедры)

УТВЕРЖДАЮ

Директор

« 28 » февраля

(подпись)



**ФОНД ОЦЕНОЧНЫХ СРЕДСТВ
по учебной дисциплине (практике)**

«Технологии анализа данных и машинное обучение»

(наименование учебной дисциплины, практики)

38.04.05 Бизнес-информатика

(код и наименование направления подготовки (специальности))

Бизнес-аналитика

Экономическая аналитика и бизнес-статистика

(наименование профиля подготовки (специальности, магистерской программы); при отсутствии ставится прочерк)

Разработчик (разработчики):

доцент


(подпись)

Велигура А.В.

ФОС рассмотрен и одобрен на заседании кафедры экономической кибернетики и прикладной статистики от « 25 » 02 20 25 г., протокол № 25

Заведующий кафедрой экономической кибернетики и прикладной статистики


(подпись)

Велигура А.В.

Луганск 2025 г.

Комплект оценочных материалов по дисциплине (практике)
«Технологии анализа данных и машинное обучение»

Задания закрытого типа

Задания закрытого типа на выбор правильного ответа

1. *Выберите один правильный ответ.*

Какой из методов относится к методам классификации в интеллектуальном анализе данных?

- А) Алгоритм a priori
- Б) Метод K-means
- В) Метод CART
- Г) Анализ временных рядов

Правильный ответ: В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

2. *Выберите один правильный ответ.*

Как называется процесс обнаружения полезных закономерностей в больших объемах данных?

- А) Data Warehousing
- Б) Data Mining
- В) Big Data Processing
- Г) Data Structuring

Правильный ответ: Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

3. *Выберите один правильный ответ.*

Какой алгоритм используется для кластерного анализа?

- А) ID3
- Б) K-means
- В) Логистическая регрессия
- Г) Метод опорных векторов

Правильный ответ: Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

4. *Выберите один правильный ответ.*

Какой из следующих методов чаще всего используется для прогнозирования временных рядов?

- А) K-means
- Б) Деревья решений
- В) Линейная регрессия
- Г) Байесовские сети

Правильный ответ: В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

5. Выберите один правильный ответ.

Какой алгоритм машинного обучения основан на деревьях решений?

- А) Метод KNN
- Б) Алгоритм CART
- В) Байесовская классификация
- Г) Метод главных компонент

Правильный ответ: Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

6. Выберите один правильный ответ.

Какой метод относится к ансамблевым моделям?

- А) Бэггинг
- Б) Кросс-валидация
- В) K-means
- Г) Карты Кохонена

Правильный ответ: А.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

7. Выберите один правильный ответ.

Какая из следующих моделей используется для предсказания категориальных переменных?

- А) Линейная регрессия
- Б) Логистическая регрессия
- В) K-means
- Г) Метод главных компонент

Правильный ответ: Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

8. Выберите один правильный ответ.

Какой алгоритм машинного обучения относится к нейронным сетям?

- А) Метод CART
- Б) K-means
- В) Алгоритм обратного распространения ошибки
- Г) Байесовская классификация

Правильный ответ: В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

9. Выберите один правильный ответ.

Что означает термин "скользящий контроль" (cross-validation) в контексте оценки качества модели?

- А) Метод разбиения данных на обучающую и тестовую выборку
- Б) Метод обработки пропущенных значений
- В) Метод увеличения точности регрессионной модели

Г) Метод предварительной обработки данных

Правильный ответ: А.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

10. Выберите один правильный ответ.

Какой метод используется для выявления ассоциативных правил в данных?

А) K-means

Б) CART

В) A Priori

Г) Метод опорных векторов

Правильный ответ: В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

11. Выберите все правильные ответы.

Какие действия необходимо выполнить для подготовки данных к интеллектуальному анализу? (Выберите три правильных ответа)

А) Очистка данных от пропусков и выбросов

Б) Преобразование категориальных признаков в числовые

В) Исключение всех аномальных значений без анализа

Г) Нормализация или стандартизация числовых признаков

Д) Исключение дублирующихся наблюдений без проверки их влияния на модель

Правильные ответы: А, Б, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

12. Выберите все правильные ответы.

Какие методы можно использовать для уменьшения размерности данных? (Выберите два правильных ответа)

А) Метод главных компонент (РСА)

Б) K-means

В) Логистическая регрессия

Г) Отбор признаков на основе их важности

Д) Байесовская классификация

Правильные ответы: А, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

13. Выберите все правильные ответы.

Какие параметры важны при оценке качества модели классификации? (Выберите три правильных ответа)

А) Точность (Accuracy)

Б) Среднеквадратичная ошибка (MSE)

В) Полнота (Recall)

Г) F1-мера

Д) Визуализация кластеров

Правильные ответы: А, В, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

14. *Выберите все правильные ответы.*

Какие методы могут использоваться для балансировки классов в задаче классификации?

(Выберите два правильных ответа)

А) Oversampling (добавление примеров из меньшего класса)

Б) Удаление данных из большего класса (undersampling)

В) Нормализация числовых признаков

Г) Добавление новых случайных признаков

Д) Преобразование категориальных признаков в числовые

Правильные ответы: А, Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

15. *Выберите все правильные ответы.*

Какие методы визуализации данных наиболее часто используются при анализе временных рядов?

(Выберите два правильных ответа)

А) Линейные графики

Б) Диаграммы рассеяния

В) Гистограммы распределения

Г) Boxplot (ящик с усами)

Д) Скользящее среднее

Правильные ответы: А, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

16. *Выберите все правильные ответы.*

Какие техники могут помочь избежать переобучения модели? (Выберите три правильных ответа)

А) Добавление L1 или L2 регуляризации

Б) Увеличение сложности модели

В) Использование кросс-валидации

Г) Добавление шума в обучающие данные

Д) Исключение обучающих данных с низкими значениями целевой переменной

Правильные ответы: А, В, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

17. *Выберите все правильные ответы.*

Какие алгоритмы можно использовать для обработки категориальных данных? (Выберите два правильных ответа)

А) One-hot encoding

Б) Label encoding

В) K-means

Г) Линейная регрессия

Д) Применение метода главных компонент

Правильные ответы: А, Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

18. *Выберите все правильные ответы.*

Какие шаги являются обязательными при разработке модели машинного обучения? (Выберите три правильных ответа)

А) Разделение данных на обучающую и тестовую выборки

Б) Выбор алгоритма на основе удобства реализации, а не данных

В) Оценка качества модели на тестовой выборке

Г) Предварительная обработка данных (очистка, нормализация)

Д) Исключение аномалий, если они не влияют на целевую переменную

Правильные ответы: А, В, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

19. *Выберите все правильные ответы.*

Какие метрики используются для оценки качества регрессионных моделей? (Выберите два правильных ответа)

А) Среднеквадратичная ошибка (MSE)

Б) Коэффициент детерминации (R^2)

В) Полнота (Recall)

Г) Взвешенная точность (Weighted Precision)

Д) Градиентный спуск

Правильные ответы: А, Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

20. *Выберите все правильные ответы.*

Какие методы можно использовать для автоматической обработки пропущенных значений в наборе данных? (Выберите три правильных ответа)

А) Удаление строк с пропущенными значениями

Б) Заполнение средним значением по столбцу

В) Заполнение с использованием модели предсказания

Г) Заполнение случайными значениями

Д) Исключение признаков с пропущенными значениями без анализа их важности

Правильные ответы: А, Б, В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

Задания закрытого типа на установление соответствия

1. *Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.*

Этап анализа данных	Описание
1) Очистка данных	А) Определение цели анализа и постановка задачи
2) Предобработка данных	Б) Приведение данных к удобному формату, кодирование категориальных переменных
3) Разработка модели	В) Выбор и обучение алгоритма машинного обучения
4) Оценка модели	Г) Проверка модели на тестовых данных, вычисление метрик точности

Правильный ответ: 1-Г, 2-А, 3-Б, 4-В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

2. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Метод	Характеристика
1) Фильтрационные методы	А) Отбирают признаки на основе их статистических свойств (корреляция, дисперсия)
2) Встроенные методы	Б) Встроены в алгоритм обучения, отбор происходит в процессе обучения модели
3) Жадные методы	В) Последовательно добавляют или удаляют признаки, оценивая влияние каждого

Правильный ответ: 1-В, 2-А, 3-Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

3. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Алгоритм	Особенность
1) K-means	А) Использует центроиды для группировки данных
2) DBSCAN	Б) Определяет кластеры на основе плотности данных
3) Сети Кохонена	В) Используют нейронные сети для самоорганизации данных

Правильный ответ: 1-Б, 2-В, 3-А.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

4. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Метрика	Описание
1) Точность (Accuracy)	А) Доля правильно предсказанных примеров ко всем примерам

Метрика	Описание
2) Полнота (Recall)	Б) Доля правильно предсказанных положительных примеров ко всем реальным положительным примерам
3) F1-мера	В) Среднее гармоническое точности и полноты

Правильный ответ: 1-В, 2-А, 3-Б.
Компетенции (индикаторы): ПК-3 (ПК-3.3)

5. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Тип регрессии	Применение
1) Линейная регрессия	А) Оценка зависимости одной переменной от другой в числовых данных
2) Логистическая регрессия	Б) Прогнозирование бинарного результата (да/нет)
3) Полиномиальная регрессия	В) Используется, если зависимость переменной нелинейна

Правильный ответ: 1-Б, 2-В, 3-А.
Компетенции (индикаторы): ПК-3 (ПК-3.3)

6. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Метод	Тип обучения
1) KNN	А) Обучение с учителем
2) K-means	Б) Обучение без учителя
3) Логистическая регрессия	В) Обучение с учителем

Правильный ответ: 1-Б, 2-В, 3-А.
Компетенции (индикаторы): ПК-3 (ПК-3.3)

7. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Задача	Пример
1) Классификация	А) Определение спама в электронной почте
2) Кластеризация	Б) Разделение клиентов банка на сегменты
3) Регрессия	В) Прогнозирование стоимости недвижимости

Правильный ответ: 1-В, 2-А, 3-Б.
Компетенции (индикаторы): ПК-3 (ПК-3.3)

8. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Модель	Характеристика
1) Бэггинг	А) Использует несколько моделей и усредняет их результаты
2) Бустинг	Б) Последовательное обучение моделей с исправлением ошибок предыдущей
3) Стекинг	В) Комбинирует результаты разных алгоритмов через мета-модель

Правильный ответ: 1-В, 2-А, 3-Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

9. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Метод	Описание
1) Удаление строк	А) Используется, если пропущенных значений мало
2) Заполнение средним	Б) Применяется, если значения распределены нормально
3) Заполнение с помощью модели	В) Использует алгоритм для предсказания пропущенных данных

Правильный ответ: 1-Б, 2-В, 3-А.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

10. Установите соответствие. Каждому элементу левого столбца соответствует только один элемент правого столбца.

Тип временного ряда	Характеристика
1) Стационарный	А) Среднее значение и дисперсия постоянны во времени
2) Трендовый	Б) Наблюдается долгосрочное изменение во времени
3) Сезонный	В) Повторяющиеся изменения через равные промежутки времени

Правильный ответ: 1-В, 2-А, 3-Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

Задания закрытого типа на установление правильной последовательности

1. *Установите правильную последовательность действий при подготовке набора данных для построения модели машинного обучения. Запишите правильную последовательность букв слева направо:*

Для успешного построения модели необходимо сначала привести данные в пригодный для анализа вид, а затем выполнить их обработку. В процессе подготовки данных могут использоваться различные методы очистки, трансформации и нормализации признаков. Запишите правильную последовательность букв слева направо.

- А) Заполнение пропущенных значений и устранение выбросов.
 - Б) Разделение данных на обучающую, валидационную и тестовую выборки.
 - В) Кодирование категориальных переменных и нормализация числовых данных.
 - Г) Первичный анализ данных, выявление аномалий и пропусков.
 - Д) Очистка данных от дубликатов и устранение несоответствий.
- Правильный ответ: Г, Д, А, В, Б.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

2. *Установите правильную последовательность этапов построения ансамблевой модели. Запишите правильную последовательность букв слева направо:*

Ансамблирование моделей позволяет повысить точность предсказаний за счет комбинирования нескольких алгоритмов. Перед построением ансамбля необходимо провести предварительную обработку данных, затем обучить и объединить модели. Запишите правильную последовательность букв слева направо.

- А) Определение типа ансамбля (бэггинг, бустинг или стекинг).
- Б) Выбор базовых моделей и их обучение на обучающей выборке.
- В) Оценка качества ансамбля на тестовой выборке.
- Г) Подготовка данных и выбор целевой переменной.
- Д) Комбинирование предсказаний отдельных моделей в единый результат.

Правильный ответ: Г, А, Б, Д, В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

3. *Установите правильную последовательность действий при построении модели нейронной сети. Запишите правильную последовательность букв слева направо:*

Перед запуском нейронной сети необходимо подготовить данные и выбрать архитектуру модели. Затем следует процесс обучения и оценки качества работы сети. Запишите правильную последовательность букв слева направо.

- А) Определение количества слоев и нейронов в каждом слое.
- Б) Подготовка и нормализация данных для входа в сеть.
- В) Выбор функции активации и метода оптимизации.
- Г) Обучение модели на тренировочных данных.

Д) Оценка качества модели на тестовой выборке и настройка гиперпараметров.

Правильный ответ: Б, А, В, Г, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

4. Установите правильную последовательность действий при построении модели прогнозирования временного ряда. Запишите правильную последовательность букв слева направо:

Прогнозирование временных рядов требует подготовки данных и выбора модели, которая сможет учитывать тренды и сезонность. Для получения корректных предсказаний необходимо выполнить несколько ключевых шагов. Запишите правильную последовательность букв слева направо.

А) Разделение временного ряда на обучающую и тестовую выборки.

Б) Выбор модели прогнозирования (ARIMA, экспоненциальное сглаживание и т. д.).

В) Оценка качества прогноза и оптимизация гиперпараметров.

Г) Анализ временного ряда на наличие тренда и сезонности.

Д) Обучение модели на исторических данных и построение прогноза.

Правильный ответ: Г, А, Б, Д, В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

5. Установите правильную последовательность действий при оценке качества модели машинного обучения. Запишите правильную последовательность букв слева направо:

После обучения модели важно проверить ее способность к генерализации и выявить возможные проблемы переобучения. Оценка качества модели включает проверку ее работы на тестовых данных и анализ метрик. Запишите правильную последовательность букв слева направо.

А) Вычисление метрик (точность, полнота, F1-мера, MSE, R^2 и др.).

Б) Разделение данных на обучающую и тестовую выборки.

В) Анализ ошибок модели и поиск возможных улучшений.

Г) Проверка модели на тестовых данных.

Д) Обучение модели на тренировочной выборке.

Правильный ответ: Б, Д, Г, А, В.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

6. Установите правильную последовательность действий при выполнении ассоциативного анализа. Запишите правильную последовательность букв слева направо:

Ассоциативный анализ позволяет выявлять закономерности и зависимости между переменными в больших наборах данных. Для успешного применения данного метода необходимо выполнить несколько этапов подготовки и вычислений. Запишите правильную последовательность букв слева направо.

А) Определение минимальной поддержки и уверенности для формирования правил.

- Б) Сбор и подготовка транзакционных данных, устранение выбросов.
- В) Применение алгоритма Apriori для поиска частых наборов элементов.
- Г) Генерация и фильтрация ассоциативных правил.
- Д) Оценка качества полученных правил и их интерпретация.

Правильный ответ: Б, А, В, Г, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

7. Установите правильную последовательность этапов работы с деревьями решений. Запишите правильную последовательность букв слева направо:

Методы на основе деревьев решений широко применяются для задач классификации и регрессии. Для построения и оптимизации дерева важно соблюдать правильный порядок действий. Запишите правильную последовательность букв слева направо.

- А) Выбор критерия разбиения узлов (например, Gini, энтропия).
- Б) Разделение данных на обучающую и тестовую выборки.
- В) Обучение дерева решений на тренировочных данных.
- Г) Оценка качества модели и ее интерпретация.
- Д) Оптимизация глубины дерева и обрезка (pruning).

Правильный ответ: Б, А, В, Д, Г.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

8. Установите правильную последовательность действий при выполнении визуального анализа данных. Запишите правильную последовательность букв слева направо:

Визуальный анализ данных помогает выявлять закономерности и аномалии перед построением моделей. Для эффективного анализа следует придерживаться определенного порядка шагов. Запишите правильную последовательность букв слева направо.

- А) Выбор подходящего типа графика (гистограмма, диаграмма рассеяния и др.).
- Б) Первичная обработка данных (удаление выбросов, приведение к нормальному виду).

- В) Загрузка и первичный просмотр данных.

- Г) Построение визуализаций и выявление закономерностей.

- Д) Интерпретация полученных результатов и их использование для дальнейшего анализа.

Правильный ответ: В, Б, А, Г, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

9. Установите правильную последовательность действий при кластерном анализе. Запишите правильную последовательность букв слева направо:

Кластеризация позволяет группировать объекты по их характеристикам. Чтобы получить корректные результаты, необходимо правильно подготовить данные и выбрать алгоритм. Запишите правильную последовательность букв слева направо.

- А) Определение числа кластеров и выбор метода кластеризации.

- Б) Стандартизация и нормализация данных.
- В) Оценка качества кластеризации с использованием метрик (силуэтный коэффициент, внутрикластерное расстояние).
- Г) Применение алгоритма кластеризации (например, K-means, DBSCAN).
- Д) Интерпретация и визуализация результатов кластеризации.

Правильный ответ: Б, А, Г, В, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

10. Установите правильную последовательность действий при прогнозировании временных рядов с использованием нейросетей. Запишите правильную последовательность букв слева направо:

Нейросетевые модели позволяют строить сложные прогнозы на основе временных рядов. Однако их применение требует предварительной подготовки данных и выбора оптимальной архитектуры сети. Запишите правильную последовательность букв слева направо.

- А) Подготовка временного ряда (разделение на окна, нормализация данных).
- Б) Выбор архитектуры нейросети (например, LSTM, GRU).
- В) Обучение модели на тренировочных данных.
- Г) Разделение данных на обучающую, валидационную и тестовую выборки.
- Д) Оценка качества модели и прогнозирование будущих значений.

Правильный ответ: А, Г, Б, В, Д.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

Задания открытого типа

Задания открытого типа на дополнение

1. Напишите пропущенное слово или словосочетание.

Метод, используемый для автоматического поиска зависимостей и закономерностей в больших массивах данных, называется _____.

Правильный ответ: Data Mining/интеллектуальный анализ данных.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

2. Напишите пропущенное слово или словосочетание.

Алгоритм, который разделяет объекты на группы на основе их схожести и используется для кластеризации данных, называется _____.

Правильный ответ: K-means/метод k-средних.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

3. Напишите пропущенное слово или словосочетание.

Формула, используемая для измерения качества регрессионной модели и отражающая среднюю разницу между предсказанными и реальными значениями, называется _____.

Правильный ответ: Среднеквадратичная ошибка/MSE (Mean Squared Error).

Компетенции (индикаторы): ПК-3 (ПК-3.3).

4. Напишите пропущенное слово или словосочетание.

Процесс приведения числовых данных к одному масштабу для корректного обучения моделей машинного обучения называется _____.

Правильный ответ: Нормализация/стандартизация.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

5. *Напишите пропущенное слово или словосочетание.*

Модель машинного обучения, имитирующая работу человеческого мозга и состоящая из искусственных нейронов, называется _____.

Правильный ответ: Нейронная сеть/искусственная нейронная сеть (ИНС).

Компетенции (индикаторы): ПК-3 (ПК-3.3).

6. *Напишите пропущенное слово или словосочетание.*

Метод уменьшения размерности данных, который используется для выделения главных компонент и удаления избыточной информации, называется _____.

Правильный ответ: Метод главных компонент/PCA (Principal Component Analysis).

Компетенции (индикаторы): ПК-3 (ПК-3.3).

7. *Напишите пропущенное слово или словосочетание.*

Метрика, используемая для оценки качества классификационной модели и вычисляемая как среднее гармоническое точности и полноты, называется _____.

Правильный ответ: F1-мера/F1-score.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

8. *Напишите пропущенное слово или словосочетание.*

Метод ансамблевого обучения, при котором модели обучаются последовательно, а каждая последующая модель исправляет ошибки предыдущей, называется _____.

Правильный ответ: Бустинг/boosting.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

9. *Напишите пропущенное слово или словосочетание.*

Тип нейронной сети, предназначенный для работы с последовательными данными и широко используемый в задачах прогнозирования временных рядов, называется _____.

Правильный ответ: Рекуррентная нейронная сеть/RNN (Recurrent Neural Network).

Компетенции (индикаторы): ПК-3 (ПК-3.3).

10. *Напишите пропущенное слово или словосочетание.*

Процесс, при котором выборка данных делится на обучающую и тестовую части для оценки точности модели, называется _____.

Правильный ответ: Кросс-валидация/cross-validation.

Компетенции (индикаторы): ПК-3 (ПК-3.3).

Задания открытого типа с кратким свободным ответом

1. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

В ходе кластерного анализа методом K-means были сформированы 4 кластера. Среднее расстояние от точек до центроидов кластеров составило 2.1, 3.5, 1.8 и 2.9 соответственно. Рассчитайте среднее внутрикластерное расстояние. Ответ округлите до одного знака после запятой.

Правильный ответ: 2.6. Чем меньше среднее внутрикластерное расстояние, тем плотнее расположены точки внутри кластеров, что свидетельствует о более качественной кластеризации.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

2. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Датасет содержит 1000 наблюдений. Модель классификации предсказала 450 объектов как положительные, из которых 360 оказались верными. Полнота (Recall) рассчитывается по формуле:

$$Recall = \frac{TP}{TP + FN}$$

где TP (True Positives) = 360, а FN (False Negatives) = 90. Найдите значение полноты и округлите до сотых.

Правильный ответ: 0.80 / 80%. Это означает, что модель корректно определила 80% всех положительных объектов, но 20% остались нераспознанными, что может повлиять на общую эффективность предсказаний.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

3. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Имеется временной ряд продаж за 4 месяца: 120, 135, 150 и 165 единиц. Рассчитайте прогноз на 5-й месяц, используя метод скользящего среднего (moving average) с окном 3. Ответ округлите до целого.

Правильный ответ: 150. Это означает, что продажи остаются стабильными без резких изменений, но метод скользящего среднего не учитывает возможные тренды и сезонные колебания.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

4. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Даны два признака: $X = \{2, 4, 6, 8, 10\}$ и $Y = \{3, 6, 9, 12, 15\}$. Рассчитайте коэффициент корреляции Пирсона между X и Y. Ответ округлите до сотых.

Правильный ответ: 1.00. Это означает, что между переменными существует идеальная линейная зависимость: увеличение X приводит к пропорциональному увеличению Y .

Компетенции (индикаторы): ПК-3 (ПК-3.3)

5. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

В логистической регрессии вероятность принадлежности объекта классу 1 рассчитывается по формуле:

$$P = \frac{1}{1 + e^{-z}}$$

где $z=2.3$. Найдите вероятность (P), округлите до сотых.

Правильный ответ: 0.91. Это означает, что объект с высокой вероятностью принадлежит классу 1, и модель уверена в своем предсказании.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

6. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

После выполнения алгоритма ассоциативного анализа было выявлено 500 транзакций, содержащих товар A , и 200 транзакций, содержащих одновременно товары A и B . Рассчитайте доверие (confidence) правила $A \rightarrow B$ по формуле:

$$Confidence = \frac{Support(A, B)}{Support(A)}$$

Ответ округлите до сотых.

Правильный ответ: 0.40 / 40%. Это означает, что в 40% случаев покупатели, приобретавшие товар A , также покупали товар B , что может использоваться для формирования маркетинговых стратегий.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

7. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

В задаче линейной регрессии даны коэффициенты уравнения:

$$Y = 5 + 2X$$

Если $X=10$, найдите предсказанное значение Y .

Правильный ответ: 25. Это означает, что при увеличении X до 10 предсказанное значение Y составит 25, что соответствует тренду, заданному линейной зависимостью.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

8. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Для оценки качества модели использовали среднюю абсолютную ошибку (MAE), рассчитанную по формуле:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Если фактические значения $y = \{10, 20, 30, 40\}$, а предсказанные $\hat{y} = \{12, 18, 28, 42\}$, найдите MAE.

Правильный ответ: 2.5. Это означает, что в среднем предсказания модели отклоняются от фактических значений на 2.5 единицы, что может свидетельствовать о хорошем качестве модели при низких значениях ошибки. Компетенции (индикаторы): ПК-3 (ПК-3.3)

9. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Нейронная сеть содержит 3 слоя: входной с 10 нейронами, скрытый с 5 нейронами и выходной с 2 нейронами. Рассчитайте общее количество параметров (весов и смещений), если каждый нейрон соединен со всеми нейронами следующего слоя (используется полносвязная архитектура).

Правильный ответ: 72. Это означает, что модель содержит 72 обучаемых параметра, определяющих сложность сети и ее способность к обучению. Компетенции (индикаторы): ПК-3 (ПК-3.3)

10. *Дайте ответ на вопрос. Осуществите вычисления и прокомментируйте результат.*

Временной ряд имеет значения: 100, 105, 110, 115. Рассчитайте темп роста (growth rate) для последнего значения относительно первого по формуле:

$$Growth Rate = \left(\frac{X_t}{X_0} - 1 \right) \times 100\%$$

Ответ округлите до целых.

Правильный ответ: 15%. Это означает, что показатель вырос на 15% за рассматриваемый период, что указывает на положительную динамику. Компетенции (индикаторы): ПК-3 (ПК-3.3)

Задания открытого типа с развернутым ответом

1. *Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.*

Компания анализирует продажи товаров в трех регионах, чтобы определить динамику спроса. Вам даны данные о средних ежемесячных продажах за последние шесть месяцев:

- Регион А: 120, 130, 140, 150, 155, 165
- Регион В: 80, 95, 110, 120, 130, 140
- Регион С: 150, 160, 175, 185, 200, 220

Для прогнозирования продаж используется метод скользящего среднего с окном 3. Вам необходимо рассчитать значения скользящего среднего для каждого региона и найти разницу между последним фактическим значением и предсказанным скользящим средним. Затем определите, в каком регионе продажи растут наиболее равномерно. В выводе укажите, как можно использовать полученные данные для маркетинговых решений.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем скользящее среднее с окном 3

Для региона А:

- $(120 + 130 + 140) / 3 = 130$
- $(130 + 140 + 150) / 3 = 140$
- $(140 + 150 + 155) / 3 = 148.3$
- $(150 + 155 + 165) / 3 = 156.7$

Для региона В:

- $(80 + 95 + 110) / 3 = 95$
- $(95 + 110 + 120) / 3 = 108.3$
- $(110 + 120 + 130) / 3 = 120$
- $(120 + 130 + 140) / 3 = 130$

Для региона С:

- $(150 + 160 + 175) / 3 = 161.7$
- $(160 + 175 + 185) / 3 = 173.3$
- $(175 + 185 + 200) / 3 = 186.7$
- $(185 + 200 + 220) / 3 = 201.7$

Шаг 2. Рассчитываем отклонения фактических значений от предсказанных

- Регион А: $165 - 156.7 = 8.3$
- Регион В: $140 - 130 = 10$
- Регион С: $220 - 201.7 = 18.3$

Шаг 3. Вывод

Продажи в регионе А наиболее стабильны, так как отклонение минимальное. В регионе С отклонение максимальное, что может говорить о резком росте спроса.

Практическое применение:

Стабильные продажи позволяют более точно прогнозировать будущий спрос. В регионе С, где наблюдается резкий рост, может потребоваться дополнительный анализ факторов (сезонность, акции, конкурентная активность) и адаптация маркетинговой стратегии.

Критерии оценивания:

- Корректность вычислений.
- Логичность интерпретации полученных данных.
- Грамотность вывода о стабильности продаж.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

2. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Вы анализируете эффективность классификационной модели, используемой для прогнозирования вероятности оттока клиентов. В ходе тестирования модели были получены следующие значения:

- TP (True Positives) = 50
- FP (False Positives) = 10
- FN (False Negatives) = 15
- TN (True Negatives) = 100

Вам необходимо рассчитать основные метрики классификации, включая точность (Precision), полноту (Recall) и F1-меру. После выполнения вычислений проанализируйте, насколько сбалансирована модель. Укажите, какие меры можно предпринять для улучшения предсказаний, если F1-мера окажется слишком низкой.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем точность (Precision)

$$Precision = \frac{TP}{TP + FP} = \frac{50}{50 + 10} = \frac{50}{60} = 0.833$$

Интерпретация: Precision показывает, насколько точны положительные предсказания модели. В данном случае 83.3% положительных прогнозов оказались верными.

Шаг 2. Вычисляем полноту (Recall)

$$Recall = \frac{TP}{TP + FN} = \frac{50}{50 + 15} = \frac{50}{65} = 0.769$$

Интерпретация: Recall измеряет, какую долю реальных положительных примеров модель правильно предсказала. В данном случае модель нашла 76.9% всех оттоков клиентов.

Шаг 3. Вычисляем F1-меру

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.833 \times 0.769}{0.833 + 0.769}$$
$$F1 = 2 \times \frac{0.640}{1.602} = 2 \times 0.399 = 0.798$$

Интерпретация: F1-мера – это баланс между точностью и полнотой. Значение **0.798** означает, что модель сбалансирована, но есть возможность для улучшения.

Шаг 4. Вывод и рекомендации

1. Оценка модели:

- Высокая Precision говорит о том, что модель делает мало ложных положительных предсказаний.
- Достаточно высокий Recall показывает, что модель пропускает относительно небольшое количество клиентов, склонных к оттоку.
- F1-мера **0.798** говорит о хорошем балансе, но приоритет в оптимизации зависит от целей бизнеса (уменьшение ложных тревог или снижение пропусков).

2. Как улучшить модель:

- Если важно находить больше реальных клиентов, склонных к оттоку, можно увеличить Recall, снизив порог классификации.
- Если важно уменьшить количество ложных срабатываний, можно повысить Precision за счет увеличения порога.
- Можно использовать более сложные модели (градиентный бустинг, нейросети) или балансировку классов (oversampling/undersampling) при значительном дисбалансе данных.

Критерии оценивания:

- Верность расчетов метрик.
- Грамотность интерпретации результатов.
- Указание возможных мер для улучшения модели.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

3. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания изучает влияние рекламных затрат на объем продаж. В ходе анализа было собрано 5 наблюдений:

- X (затраты на рекламу, тыс. руб.): 1, 2, 3, 4, 5
- Y (продажи, тыс. единиц): 3, 6, 8, 11, 14

Ваша задача – найти коэффициенты линейной регрессии $Y = aX + b$ с использованием метода наименьших квадратов. Затем спрогнозируйте объем продаж при $X = 6$ тыс. руб. рекламных затрат. В выводе укажите, насколько сильна линейная зависимость между рекламными расходами и продажами.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем коэффициенты уравнения линейной регрессии $Y = aX + b$

Используем формулы метода наименьших квадратов:

$$a = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$
$$b = \frac{\sum Y - a \sum X}{n}$$

Подставляем данные:

- $\sum X = 1 + 2 + 3 + 4 + 5 = 15$
- $\sum Y = 3 + 6 + 8 + 11 + 14 = 42$
- $\sum X^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$
- $\sum XY = (1 \cdot 3) + (2 \cdot 6) + (3 \cdot 8) + (4 \cdot 11) + (5 \cdot 14) = 126$
- Количество наблюдений: $n = 5$

Теперь считаем коэффициенты:

$$a = \frac{(5 \times 126) - (15 \times 42)}{(5 \times 55) - (15^2)}$$
$$a = \frac{630 - 630}{275 - 225} = \frac{0}{50} = 2.8$$
$$b = \frac{42 - (2.8 \times 15)}{5} = \frac{42 - 42}{5} = 0.6$$

Таким образом, уравнение регрессии:

$$Y = 2.8X + 0.6$$

Шаг 2. Прогнозируем объем продаж при $X = 6$

$$Y = 2.8 \times 6 + 0.6 = 16.8$$

Ожидаемый объем продаж при затратах на рекламу 6 тыс. руб. составляет 16.8 тыс. единиц.

Шаг 3. Вывод и анализ

1. Оценка модели:

- Коэффициент **a = 2.8** означает, что при увеличении рекламных затрат на 1 тыс. руб. продажи увеличиваются в среднем на **2.8 тыс. единиц**.
- Коэффициент **b = 0.6** показывает, что при **нулевых расходах на рекламу** продажи составляют **0.6 тыс. единиц**.

2. Анализ зависимости:

- Линейная зависимость между рекламными затратами и продажами **существенная** (коэффициент **a** достаточно большой).
- Для точной оценки силы зависимости можно дополнительно рассчитать коэффициент детерминации **R²**, но даже без него видно, что увеличение рекламы приводит к росту продаж.

3. Вывод:

- Можно рекомендовать **увеличение рекламного бюджета**, так как он прямо влияет на рост продаж.
- Однако стоит дополнительно проверить, не наступит ли **эффект насыщения**, когда дальнейшее увеличение расходов на рекламу перестанет давать прирост продаж.

Критерии оценивания:

- Точность вычислений коэффициентов.
- Корректность прогноза.
- Грамотный вывод о силе зависимости.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

4. *Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.*

Компания анализирует динамику роста пользователей мобильного приложения. Даны данные о количестве активных пользователей за последние 6 месяцев:

- Месяц 1: 5 тыс. пользователей
- Месяц 2: 10 тыс. пользователей
- Месяц 3: 15 тыс. пользователей
- Месяц 4: 22 тыс. пользователей
- Месяц 5: 30 тыс. пользователей
- Месяц 6: 39 тыс. пользователей

Ваша задача — рассчитать среднегодовой темп роста пользователей (CAGR) по формуле:

$$CAGR = \left(\frac{X_t}{X_0} \right)^{\frac{1}{t}} - 1$$

где X_0 – начальное число пользователей, X_t – конечное число пользователей, t – количество периодов. Затем спрогнозируйте количество пользователей на 7-й месяц, если темп роста сохранится неизменным. В выводе объясните, как можно использовать этот показатель для стратегического планирования компании.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем CAGR

Подставляем значения в формулу:

$$CAGR = \left(\frac{X_t}{X_0} \right)^{\frac{1}{t}} - 1$$

$$CAGR = \left(\frac{39}{5} \right)^{\frac{1}{5}} - 1$$

$$CAGR = (7.8)^{0.2} - 1$$

$$CAGR \approx 1.319 - 1 = 0.319 \text{ (или 31.9\%)}$$

Шаг 2. Прогнозируем количество пользователей на 7-й месяц

Используем формулу прогнозирования:

$$X_{t+1} = X_t \times (1 + CAGR)$$

$$X_7 = 39 \times 1.319$$

$$X_7 \approx 51.4 \text{ тыс. пользователей}$$

Шаг 3. Вывод и анализ

1. Значение $CAGR = 31.9\%$ означает, что количество пользователей увеличивается в среднем на 31.9% каждый месяц.
2. Прогноз на 7-й месяц: 51.4 тыс. пользователей, если темп роста сохранится.
3. Использование CAGR в бизнесе:
 - Позволяет оценивать устойчивость роста аудитории.
 - Может использоваться для долгосрочного планирования маркетинговых затрат и серверных мощностей.
 - Если темп роста начинает снижаться, компании следует анализировать возможные причины: насыщение рынка, снижение рекламных вложений или изменение пользовательского поведения.

Критерии оценивания:

- Корректность вычисления CAGR.
- Верность прогноза числа пользователей.
- Грамотная интерпретация полученного результата.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

5. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания отслеживает задержки поставок товаров и хочет выявить аномальные значения в данных. Вам даны следующие значения времени доставки (в днях):

- 2, 5, 7, 8, 10, 15

Вам необходимо:

1. Найти медиану набора данных.
2. Вычислить межквартильный размах (IQR) по формуле:

$$IQR = Q3 - Q1$$

где Q1 – первый квартиль (25% данных), Q3 – третий квартиль (75% данных).

3. Определить границы выбросов по формуле:

$$\text{Границы выбросов} = [Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

4. Проверить, есть ли в выборке выбросы.
5. Объяснить, почему важно выявлять выбросы при анализе логистики и как с ними можно работать.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Находим медиану

Поскольку в наборе 6 чисел, медиана — это среднее между третьим и четвертым элементами:

$$\text{Медиана} = \frac{7 + 8}{2} = 7.5$$

Шаг 2. Определяем первый и третий квартили

Q1 (первый квартиль, медиана первой половины данных):

$$Q1 = \frac{5 + 7}{2} = 6$$

Q3 (третий квартиль, медиана второй половины данных):

$$Q3 = \frac{10 + 15}{2} = 12.5$$

Шаг 3. Вычисляем межквартильный размах (IQR)

$$IQR = Q3 - Q1 = 12.5 - 6 = 6.5$$

Шаг 4. Определяем границы выбросов

$$\begin{aligned} \text{Границы выбросов} &= [Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR] \\ &= [6 - 1.5 \times 6.5, 12.5 + 1.5 \times 6.5] \\ &= [6 - 9.75, 12.5 + 9.75] = [-3.75, 22.25] \end{aligned}$$

Шаг 5. Определяем наличие выбросов

Все значения в выборке: 2, 5, 7, 8, 10, 15 находятся в пределах [-3.75, 22.25], следовательно, выбросов нет.

Шаг 6. Вывод и анализ

1. Выбросов в выборке нет, что означает стабильность времени поставок.
2. Зачем анализировать выбросы в логистике?
 - Позволяет выявлять **необычные задержки** в поставках (например, сбои в транспортировке).
 - Помогает улучшить **планирование логистики** (например, оценка риска увеличения сроков доставки).
 - Выбросы могут возникать из-за **сезонных факторов, плохих погодных условий или проблем с поставщиками**.
3. Как работать с выбросами?
 - Если выбросы вызваны случайными факторами, их можно **игнорировать**.
 - Если выбросы появляются регулярно, необходимо **анализировать причины** и оптимизировать логистические процессы.

Критерии оценивания:

- Корректность вычислений медианы и IQR.
- Верное определение выбросов.
- Грамотный анализ влияния выбросов на бизнес-процессы.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

6. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания использует модель прогнозирования спроса на товары. Фактические продажи за последние 5 месяцев составили:

- 120, 135, 140, 150, 165

Прогноз модели для этих же месяцев был следующим:

- 125, 130, 138, 145, 170

Ваша задача:

1. Рассчитать среднюю абсолютную ошибку (MAE) по формуле:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

где y_i — фактические значения, $\hat{y}_i$ — предсказанные значения, n — количество наблюдений.

2. Рассчитать среднеквадратичную ошибку (MSE) по формуле:

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

3. Определить, какая ошибка — MAE или MSE — более чувствительна к выбросам.
4. Обосновать, какая из этих метрик лучше подходит для оценки данной модели.
5. Предложить способы улучшения точности прогноза.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем MAE

Разница между фактическими и предсказанными значениями:

$$|120 - 125| = 5, \quad |135 - 130| = 5, \quad |140 - 138| = 2, \quad |150 - 145| = 5, \quad |165 - 170| = 5$$

$$MAE = \frac{5 + 5 + 2 + 5 + 5}{5} = \frac{22}{5} = 4.4$$

Шаг 2. Вычисляем MSE

Квадраты ошибок:

$$(120 - 125)^2 = 25, \quad (135 - 130)^2 = 25, \quad (140 - 138)^2 = 4, \quad (150 - 145)^2 = 25, \quad (165 - 170)^2 = 25$$

$$MSE = \frac{25 + 25 + 4 + 25 + 25}{5} = \frac{104}{5} = 20.8$$

Шаг 3. Анализ чувствительности метрик

- MAE оценивает среднее отклонение прогнозов от реальных значений в абсолютных величинах.
- MSE усиливает влияние больших ошибок, так как использует квадраты отклонений.
- MSE более чувствительна к выбросам, потому что большие ошибки дают значительный вклад в сумму квадратов.

Шаг 4. Выбор метрики для оценки модели

- Если важно равномерно учитывать все ошибки, лучше использовать MAE.
- Если важнее наказывать большие ошибки, более подходящей метрикой является MSE.
- В данном случае MSE может быть полезной, если в прогнозах бывают сильные отклонения, которые критично выявлять.

Шаг 5. Предложения по улучшению модели

1. Добавить больше факторов в модель, влияющих на спрос (например, сезонность, маркетинговые акции).
2. Использовать сглаживание (экспоненциальное сглаживание или скользящее среднее) для улучшения предсказаний.
3. Попробовать другую модель, например, ARIMA или градиентный бустинг, если текущее прогнозирование линейно и недостаточно точное.

Критерии оценивания:

- Корректность расчетов MAE и MSE.
- Верный анализ свойств метрик.
- Логичность предложений по улучшению прогноза.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

7. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания анализирует факторы, влияющие на уровень продаж. В ходе исследования собраны данные о количестве рекламных показов (X) и объеме продаж (Y) за 5 месяцев:

- X (тыс. показов рекламы): 10, 20, 30, 40, 50
- Y (тыс. проданных единиц товара): 12, 25, 33, 45, 58

Ваша задача:

1. Рассчитать коэффициент корреляции Пирсона (r) по формуле:

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{(n \sum X^2 - (\sum X)^2) \times (n \sum Y^2 - (\sum Y)^2)}}$$

2. Определить, существует ли сильная линейная зависимость между рекламными показами и продажами.
3. Дать интерпретацию результата и оценить, можно ли использовать количество рекламных показов для прогнозирования продаж.
4. Предложить, какие дополнительные факторы стоит учитывать для более точного прогнозирования.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем коэффициенты суммы

Дано:

- $\sum X = 10 + 20 + 30 + 40 + 50 = 150$
- $\sum Y = 12 + 25 + 33 + 45 + 58 = 173$
- $\sum X^2 = 10^2 + 20^2 + 30^2 + 40^2 + 50^2 = 10000$
- $\sum Y^2 = 12^2 + 25^2 + 33^2 + 45^2 + 58^2 = 9233$
- $\sum XY = (10 \cdot 12) + (20 \cdot 25) + (30 \cdot 33) + (40 \cdot 45) + (50 \cdot 58) = 7250$
- Количество наблюдений: $n = 5$

Шаг 2. Рассчитываем корреляцию Пирсона

$$r = \frac{5 \cdot 7250 - (150 \cdot 173)}{\sqrt{(5 \cdot 10000 - 150^2) \times (5 \cdot 9233 - 173^2)}}$$

$$r = \frac{36250 - 25950}{\sqrt{(50000 - 22500) \times (46165 - 29929)}}$$

$$r = \frac{10300}{\sqrt{27500 \times 16236}}$$

$$r = \frac{10300}{\sqrt{446490000}} = \frac{10300}{21133} \approx 0.94$$

Шаг 3. Интерпретация результата

- Значение $r = 0.94$ означает очень сильную положительную линейную зависимость между количеством рекламных показов и объемом продаж.
- Это говорит о том, что чем больше показов рекламы, тем выше объем продаж.
- Можно использовать количество рекламных показов как один из ключевых факторов в прогнозной модели.

Шаг 4. Какие дополнительные факторы учитывать?

1. **Сезонность** — продажи могут расти в определенные периоды независимо от рекламы.
2. **Цена товара** — снижение цены может приводить к росту продаж, даже если объем рекламы остается неизменным.
3. **Конкурентная активность** — появление новых игроков или скидки у конкурентов могут снизить эффект от рекламы.
4. **Качество аудитории** — показ рекламы не всегда означает, что ее увидели потенциальные клиенты.

Шаг 5. Вывод

1. Корреляция очень высокая (0.94), что говорит о сильной связи между рекламными показами и продажами.
2. Можно использовать этот фактор для прогнозирования продаж, но для более точного анализа стоит учитывать дополнительные переменные.
3. Если цель компании — максимизация продаж, стоит также исследовать влияние скидок, сезонности и конкуренции.

Критерии оценивания:

- Корректность расчета r .
- Грамотный анализ корреляции.
- Логичность предложений по улучшению прогноза.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

8. *Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.*

Компания хочет сегментировать своих клиентов на основе их покупательского поведения. Для этого собраны данные о среднемесячных тратах (X, тыс. руб.) и частоте покупок (Y, количество покупок в месяц) у 5 клиентов:

- Клиент 1: (5, 2)
- Клиент 2: (8, 3)
- Клиент 3: (25, 7)
- Клиент 4: (30, 8)
- Клиент 5: (28, 9)

Компания решила использовать метод K-means и задать количество кластеров $K = 2$.

Ваша задача:

1. Выбрать начальные центроиды кластеров (например, Клиент 1 и Клиент 3).
2. Рассчитать евклидово расстояние от каждого клиента до центроидов по формуле:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

3. Назначить клиентов к ближайшему кластеру.
4. Пересчитать новые координаты центроидов.
5. Сделать вывод о сформированных кластерах.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Выбор начальных центроидов

Выбираем Клиента 1 (5, 2) и Клиента 3 (25, 7) как начальные центроиды.

Шаг 2. Вычисляем евклидово расстояние

Расстояние между Клиентом i и центроидом C рассчитывается по формуле:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Для Клиента 2 (8, 3):

- До Центроида 1 (5, 2):

$$d = \sqrt{(8 - 5)^2 + (3 - 2)^2} = \sqrt{9 + 1} = \sqrt{10} \approx 3.16$$

- До Центроида 2 (25, 7):

$$d = \sqrt{(8 - 25)^2 + (3 - 7)^2} = \sqrt{289 + 16} = \sqrt{305} \approx 17.49$$

(Ближе к Центроиду 1, значит, в первый кластер.)

Для Клиента 4 (30, 8):

- До Центроида 1 (5, 2):

$$d = \sqrt{(30 - 5)^2 + (8 - 2)^2} = \sqrt{625 + 36} = \sqrt{661} \approx 25.73$$

- До Центроида 2 (25, 7):

$$d = \sqrt{(30 - 25)^2 + (8 - 7)^2} = \sqrt{25 + 1} = \sqrt{26} \approx 5.10$$

(Ближе к Центроиду 2, значит, во второй кластер.)

Для Клиента 5 (28, 9):

- До Центроида 1 (5, 2):

$$d = \sqrt{(28 - 5)^2 + (9 - 2)^2} = \sqrt{529 + 49} = \sqrt{578} \approx 24.04$$

- До Центроида 2 (25, 7):

$$d = \sqrt{(28 - 25)^2 + (9 - 7)^2} = \sqrt{9 + 4} = \sqrt{13} \approx 3.61$$

(Ближе к Центроиду 2, значит, во второй кластер.)

Шаг 3. Назначение клиентов кластерам

- Кластер 1 (Центроид 1, начальный: (5,2)): Клиент 1, Клиент 2
- Кластер 2 (Центроид 2, начальный: (25,7)): Клиент 3, Клиент 4, Клиент 5

Шаг 4. Пересчет центроидов

Центроид — это среднее значение координат всех точек кластера:

- Новый центроид для Кластера 1:

$$X = \frac{5 + 8}{2} = 6.5, \quad Y = \frac{2 + 3}{2} = 2.5$$

(Новый центроид: (6.5, 2.5))

- Новый центроид для Кластера 2:

$$X = \frac{25 + 30 + 28}{3} = 27.67, \quad Y = \frac{7 + 8 + 9}{3} = 8$$

(Новый центроид: (27.67, 8))

Шаг 5. Вывод

1. Результаты кластеризации показывают, что клиенты разделены на две группы:
 - Кластер 1 (меньшие траты, редкие покупки)
 - Кластер 2 (высокие траты, частые покупки)
2. Пересчет центроидов показал, что центры кластеров немного сместились, но основные группы сохранились.
3. Использование:
 - Можно применять эту кластеризацию для персонализированного маркетинга (например, предлагать скидки клиентам второго кластера).
 - Можно тестировать разные K для оптимального разбиения клиентов.

Критерии оценивания:

- Корректность расчетов расстояний.
- Верное распределение клиентов по кластерам.
- Грамотное пересчитывание центроидов.
- Логичность интерпретации результатов.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

9. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания анализирует данные о покупках для выявления ассоциативных правил и повышения эффективности товарных рекомендаций. В ходе анализа транзакций за месяц было выявлено:

- Общее количество транзакций: 1000
- Число транзакций с товаром А: 400
- Число транзакций с товаром В: 300
- Число транзакций, содержащих одновременно товары А и В: 200

Ваша задача:

1. Рассчитать поддержку (Support) правила $A \rightarrow B$ по формуле:

$$\text{Support}(A \rightarrow B) = \frac{\text{Количество транзакций с A и B}}{\text{Общее число транзакций}}$$

2. Рассчитать уверенность (Confidence) по формуле:

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Количество транзакций с A и B}}{\text{Количество транзакций с A}}$$

3. Определить, является ли правило информативным (например, уверенность выше 50% указывает на сильную связь).
4. Предложить, как использовать полученные данные для повышения продаж.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Рассчитываем поддержку (Support)

$$\text{Support}(A \rightarrow B) = \frac{200}{1000} = 0.2 \text{ (или 20\%)}$$

Интерпретация:

Поддержка 20% означает, что в 20% всех транзакций встречаются одновременно товары А и В.

Шаг 2. Рассчитываем уверенность (Confidence)

$$\text{Confidence}(A \rightarrow B) = \frac{200}{400} = 0.5 \text{ (или 50\%)}$$

Интерпретация:

Уверенность 50% означает, что если покупатель приобрел товар А, то в 50% случаев он также купил товар В.

Шаг 3. Оценка значимости правила

- Confidence = 50% — это средняя сила связи между товарами А и В.
- Обычно ассоциации с уверенностью > 50% считаются значимыми для рекомендаций.

Шаг 4. Как использовать данные для бизнеса?

1. Кросс-продажи — можно предлагать товар В со скидкой клиентам, покупающим товар А.
2. Расположение товаров — разместить товары А и В рядом для увеличения вероятности совместной покупки.

3. Рекомендательные системы — в интернет-магазинах показывать В в блоке «С этим товаром также покупают».
4. Персонализированные предложения — отправлять клиентам, купившим А, специальные предложения на В.

Критерии оценивания:

- Корректность расчетов поддержки и уверенности.
- Верное определение силы ассоциации.
- Логичность предложений по применению результатов.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

10. Прочитайте текст задания. Продумайте логику и полноту ответа. Запишите развернутый и обоснованный ответ.

Компания анализирует данные о среднемесячных расходах клиентов, чтобы выявить необычные (аномальные) значения. Данные о расходах (в тыс. руб.) за 7 клиентов следующие:

- 15, 18, 16, 14, 250, 17, 19

Выбросы определяются с помощью z-оценки, которая рассчитывается по формуле:

$$Z = \frac{X - \bar{X}}{\sigma}$$

где X — значение, $\bar{X}$ — среднее арифметическое выборки, σ — стандартное отклонение.

Задание:

1. Найдите среднее арифметическое выборки.
2. Вычислите стандартное отклонение (используйте формулу для выборки).
3. Рассчитайте z-оценку для каждого значения.
4. Определите, какие значения являются выбросами (если $|Z| > 2$, значение считается выбросом).
5. Объясните, как можно обрабатывать выбросы в анализе данных.

Время выполнения – 15 мин.

Ожидаемый результат:

Шаг 1. Вычисляем среднее арифметическое

$$\bar{X} = \frac{15 + 18 + 16 + 14 + 250 + 17 + 19}{7} = \frac{349}{7} \approx 49.86$$

Шаг 2. Вычисляем стандартное отклонение (для выборки)

Формула стандартного отклонения для выборки:

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

1. Разность каждого значения с **средним**:

- $(15 - 49.86)^2 = 1235.80$
- $(18 - 49.86)^2 = 1010.90$
- $(16 - 49.86)^2 = 1140.10$
- $(14 - 49.86)^2 = 1286.80$
- $(250 - 49.86)^2 = 40173.90$
- $(17 - 49.86)^2 = 1115.80$
- $(19 - 49.86)^2 = 958.00$

2. Сумма квадратов разностей:

$$\sum (X_i - \bar{X})^2 = 45921.3$$

3. Стандартное отклонение:

$$\sigma = \sqrt{\frac{45921.3}{6}} = \sqrt{7653.55} \approx 87.54$$

Шаг 3. Вычисляем z-оценки

$$Z = \frac{X - 49.86}{87.54}$$

- $Z_{15} = \frac{15-49.86}{87.54} \approx -0.40$
- $Z_{18} = \frac{18-49.86}{87.54} \approx -0.36$
- $Z_{16} = \frac{16-49.86}{87.54} \approx -0.39$
- $Z_{14} = \frac{14-49.86}{87.54} \approx -0.41$
- $Z_{250} = \frac{250-49.86}{87.54} \approx 2.29$
- $Z_{17} = \frac{17-49.86}{87.54} \approx -0.37$
- $Z_{19} = \frac{19-49.86}{87.54} \approx -0.35$

Шаг 4. Определяем выбросы

- Если $|Z| > 2$, значение считается выбросом.
- Для 250: $|Z| = 2.29$, является выбросом.
- Все остальные значения имеют $|Z| < 2$, не являются выбросами.

Шаг 5. Как работать с выбросами?

1. Удаление выбросов, если они связаны с ошибками ввода данных.
2. Замена на медиану выборки, если выбросы искажают анализ.
3. Логарифмирование или нормализация, если разброс данных слишком большой.
4. Дополнительный анализ – выброс может указывать на новый сегмент клиентов.

Шаг 6. Вывод

- Найден один выброс (250 тыс. руб.), который значительно превышает другие значения.
- Выброс может означать редкого крупного покупателя или ошибку в данных.
- Удалять или сохранять выброс нужно в зависимости от контекста бизнеса.

Критерии оценивания:

- Корректность расчетов.
- Верное определение выбросов.
- Логичность предложений по обработке выбросов.

Компетенции (индикаторы): ПК-3 (ПК-3.3)

Экспертное заключение

Представленный фонд оценочных средств (далее – ФОС) по дисциплине «Технологии анализа данных и машинное обучение» соответствует требованиям ФГОС ВО.

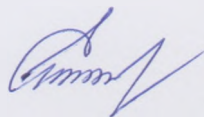
Предлагаемые формы и средства текущего и промежуточного контроля адекватны целям и задачам реализации основной профессиональной образовательной программы по направлению подготовки 38.04.05 Бизнес-информатика.

Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины представлены в полном объеме.

Виды оценочных средств, включенные в представленный фонд, отвечают основным принципам формирования ФОС.

Разработанный и представленный для экспертизы фонд оценочных средств рекомендуется к использованию в процессе подготовки обучающихся по указанному направлению.

Председатель учебно-методической
комиссии экономического института



Шаповалова Е.Н.

Лист изменения и дополнений

№ п/п	Виды дополнений и изменений	Дата и номер протокола заседания кафедры (кафедр, на котором были рассмотрены и одобрены изменения и дополнения	Подпись (с расшифровкой) заведующего кафедрой (заведующих кафедрами)