Федеральное агентство по рыболовству
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Калининградский государственный технический университет»
(ФГБОУ ВО «КГТУ»)

УТВЕРЖДАЮ
Начальник УРОПСП

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
ФГБОУ ВО «КГТУ»
СВЕДЕНИЯ О СЕРТИФИКАТЕ ЭП
Сертификат:
0174a99400e2af6fb74dca3b394e65efc4
Владелец: Мельникова Виктория Александровна
Действителен: с 12.04.2023 по 12.04.2024

Фонд оценочных средств
(приложение к рабочей программе дисциплины)

«ТЕХНОЛОГИИ DATA MINING»

основной профессиональной образовательной программы магистратуры
по направлению подготовки

09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ИНСТИТУТ цифровых технологий

РАЗРАБОТЧИК кафедра прикладной математики и информационных технологий

1 РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Таблица 1 – Планируемые результаты обучения по дисциплине, соотнесенные с установленными индикаторами достижения компетенций

Код и наименование компетенции	Индикаторы достижения компетенции	Дисциплина	Результаты обучения (владения, умения и знания), соотнесенные с компетенциями/индикаторами достижения компетенции
ПК-1: Разработка и внедрение новых методов и технологий исследования больших данных	ПК-1.2: Проведение испытаний и разработка рекомендаций по внедрению и использованию усовершенствованных или разработанных новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с большими данными	Технологии Data Mining	*Знать:* - содержание технологии Data Mining; - основные методы Data Mining; *Уметь:* - понимать основные проблемы, возникающие при анализе больших данных, и пути их решения в рамках технологии Data Mining; - выбирать методы интеллектуального анализа исходя из практической задачи; *Владеть:* - навыками анализа данных различной природы с использованием современных инструментальных средств.

2 ПЕРЕЧЕНЬ ОЦЕНОЧНЫХ СРЕДСТВ И КРИТЕРИИ ОЦЕНИВАНИЯ

2.1 К оценочным средствам текущего контроля успеваемости относятся:

- тестовые задания открытого и закрытого типов.

2.2 Промежуточная аттестация по дисциплине проводится в форме зачета, который выставляется по результатам прохождения всех видов текущего контроля успеваемости. При необходимости тестовые задания закрытого и открытого типов могут быть использованы для проведения промежуточной аттестации.

2.3 Критерии оценки результатов освоения дисциплины

Универсальная система оценивания результатов обучения включает в себя системы оценок: 1) «отлично», «хорошо», «удовлетворительно», «неудовлетворительно»; 2) «зачтено», «не зачтено»; 3) 100 – балльную/процентную систему и правило перевода оценок в пятибалльную систему (табл. 2).

Таблица 2 – Система оценок и критерии выставления оценки

Система оценок / Критерий	**2**	**3**	**4**	**5**
	0-40%	**41-60%**	**61-80 %**	**81-100 %**
	«неудовлетворительно»	**«удовлетворительно»**	**«хорошо»**	**«отлично»**
	«не зачтено»	**«зачтено»**		
1 Системность и полнота знаний в отношении изучаемых объектов	Обладает частичными и разрозненными знаниями, которые не может научно-корректно связывать между собой (только некоторые из которых может связывать между собой)	Обладает минимальным набором знаний, необходимым для системного взгляда на изучаемый объект	Обладает набором знаний, достаточным для системного взгляда на изучаемый объект	Обладает полнотой знаний и системным взглядом на изучаемый объект
2 Работа с информацией	Не в состоянии находить необходимую информацию, либо в состоянии находить отдельные фрагменты информации в рамках поставленной задачи	Может найти необходимую информацию в рамках поставленной задачи	Может найти, интерпретировать и систематизировать необходимую информацию в рамках поставленной задачи	Может найти, систематизировать необходимую информацию, а также выявить новые, дополнительные источники информации в рамках поставленной задачи
3.Научное осмысление изучаемого явления, процесса, объекта	Не может делать научно корректных выводов из имеющихся у него сведений, в состоянии проанализировать только некоторые из имеющихся у него сведений	В состоянии осуществлять научно корректный анализ предоставленной информации	В состоянии осуществлять систематический и научно корректный анализ предоставленной информации, вовлекает в исследование новые релевантные задаче данные	В состоянии осуществлять систематический и научно-корректный анализ предоставленной информации, вовлекает в исследование новые релевантные поставленной задаче данные, предлагает новые ракурсы поставленной задачи

Система оценок / Критерий	2	3	4	5
	0-40%	**41-60%**	**61-80 %**	**81-100 %**
	«неудовлетворительно»	**«удовлетворительно»**	**«хорошо»**	**«отлично»**
	«не зачтено»	**«зачтено»**		
4. Освоение стандартных алгоритмов решения профессиональных задач	В состоянии решать только фрагменты поставленной задачи в соответствии с заданным алгоритмом, не освоил предложенный алгоритм, допускает ошибки	В состоянии решать поставленные задачи в соответствии с заданным алгоритмом	В состоянии решать поставленные задачи в соответствии с заданным алгоритмом, понимает основы предложенного алгоритма	Не только владеет алгоритмом и понимает его основы, но и предлагает новые решения в рамках поставленной задачи

3 ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

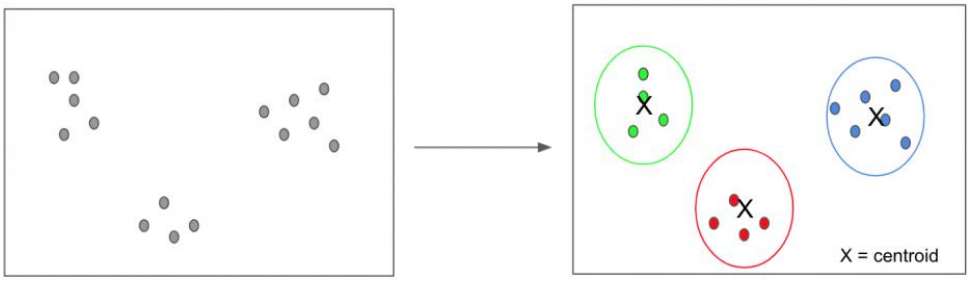
ПК-1: Разработка и внедрение новых методов и технологий исследования больших данных.

Индикатор ПК-1.2: Проведение испытаний и разработка рекомендаций по внедрению и использованию усовершенствованных или разработанных новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с большими данными.

Тестовые задания открытого типа:

1. В основе алгоритма _____ лежит построение (обучение) бинарных деревьев решений.
Введите название заглавными буквами (аббревиатура, англ.) Допускаются иные формулировки ответа, не искажающие его смысла.
Ответ: CART

2. На рисунке



представлена иллюстрация алгоритма _____
Введите общепринятое название алгоритма (англ.) Допускаются иные формулировки ответа, не искажающие его смысла.
Ответ: k-means

3. Нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке - это _____ .
Введите название явления в именительном падеже, регистр не важен.
Ответ: переобучение

4. Евклидово расстояние является частным случаем метрики Миньковского $D(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$ при значении p, равном ___ .
Введите число.
Ответ: 2

5. Торговая компания хочет на основе имеющихся у нее анкетных данных получить группы клиентов, обладающих общими характеристиками. В технологии Data Mining для решения применяется тип задачи:_____ .
Определите название в именительном падеже, регистр не важен.
Ответ: кластеризация

6. Доступны данные каталога землетрясений, содержащие сведения о дате, времени, месте, магнитуде и пр. Необходимо спрогнозировать магнитуду следующего землетрясения. В технологии Data Mining для рещения применяется тип задачи:_____ .
Определите название в именительном падеже, регистр не важен..
Ответ: регрессия

7. В технологии Data Mining для фильтрации спам-писем применяется тип задачи:_____ .
Определите название в именительном падеже, регистр не важен..
Ответ: классификация

8. В теории нейронных сетей функция _____ определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения.
Допускаются иные формулировки ответа, не искажающие его смысла.
Ответ: активации

9. В признаковом пространстве объекты, существенно отличающиеся от большинства остальных - это _____ .
Ответ: выброс

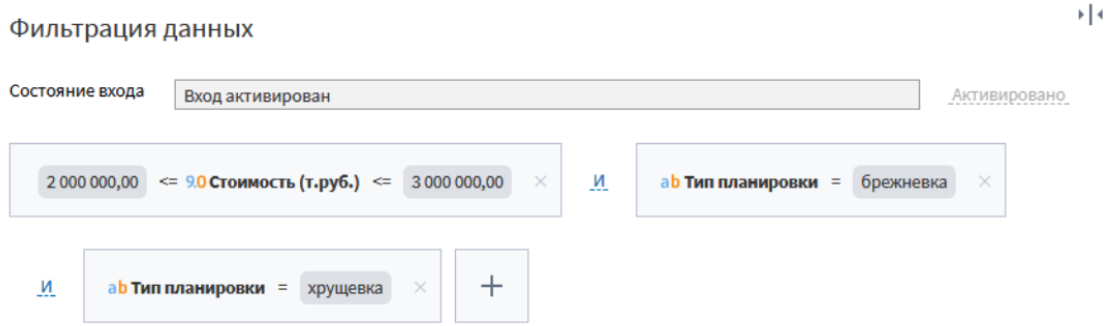
10. Имеется набор данных:

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4
6	4.6	3.4	1.4	0.3
7	5.0	3.4	1.5	0.2
8	4.4	2.9	1.4	0.2
9	4.9	3.1	1.5	0.1

При использовании системы Loginom в задаче кластеризации алгоритмом k-means для поля «sepal_length» (длина лепестка) наиболее подходящим будет _____ тип данных.
Введите название типа. Допускаются иные формулировки ответа, не искажающие его смысла.
Ответ: вещественный

Тестовые задания открытого типа (на дополнение):

1. В сценарии Loginom осуществляется фильтрация набора данных, содержащего сведения о рыночной стоимости квартир. Узел «фильтрация» настроен, как указано на рисунке.



Для корректного набора данных условиям фильтра будет удовлетворять количество записей: ____.
Введите число.
Ответ: 0

2. В системе Loginom с помощью узла «Слияние» выполнено правое соединения (RIGHT JOIN) Таблицы 1 и Таблицы 2 по полю Артикул с главной таблицей «Таблица 1».

Таблица 1

Клиент	Артикул	Кол-во
Клиент 1	1234	1
Клиент 1	2475	3
Клиент 2	6488	2
Клиент 3	1275	1

Таблица 2

Артикул	Товар	Цена
1234	Молоко	80
2475	Хлеб	30
1275	Йогурт	35
4168	Кофе	150

.

Итоговая таблица будет иметь пустые значения в: _____.
Введите имя поля (имена полей через запятую, без пробелов, регистр не важен)
Ответ: Клиент, Кол-во

3. Для решения задач бинарной классификации, в которых выходная переменная может принимать только два значения, используется модель регрессии: _____.
Введите вид модели в именительном падеже, регистр не важен.
Ответ: логистическая

Тестовые задания открытого типа (с развернутым ответом):

1. Компания решила провести анализ клиентской базы с целью оптимизации маркетинговых стратегий. Однако данные, собранные за последний год, содержат различные ошибки и неточности. Вам необходимо провести этап очистки данных, чтобы обеспечить корректность и надежность анализа. Какие шаги вы предпримете на этапе очистки данных для обработки ошибок и неточностей в клиентской базе? Укажите основные проблемы, которые могут возникнуть, и опишите методы и подходы к их решению.

Правильный ответ:
Идентификация проблем: Ошибки и неточности могут включать в себя дубликаты клиентских записей, пропущенные значения в обязательных полях, аномальные выбросы в числовых данных, а также некорректные или несогласованные значения в категориальных признаках (например, пол или возраст).

Методы исправления:

- Удаление дубликатов: Идентификация дубликатов на основе уникальных идентификаторов клиентов или других релевантных признаков и их удаление.
- Заполнение пропущенных значений: Использование методов, таких как заполнение средним значением или медианой для числовых данных и заполнение модой для категориальных данных.
- Коррекция аномалий: Идентификация и удаление или коррекция выбросов в числовых данных, если они приводят к некорректным результатам.

- Стандартизация и проверка категориальных значений: Приведение категориальных значений к единому стандарту, например, использование нижнего или верхнего регистра, и проверка наличия некорректных значений.

Критерии оценки:

Зачтено: Правильный ответ предоставляет конкретные методы для решения различных проблем. Перечислено не менее 2 проблем и предложено не менее 1 метода устранения указанных проблем.

Не зачтено: Перечислено менее 2 проблем и/или указаны методы коррекции, не относящиеся к устранению обозначенным проблемам.

2. Вы занимаетесь анализом данных по клиентам онлайн-магазина с целью определения эффективности новой рекламной кампании. Одним из важных аспектов является оценка разнородности классов - то есть, насколько хорошо новая кампания привлекла разные группы клиентов. Вам предложили использовать индекс Джини в этом контексте.

Как вы будете использовать индекс Джини для оценки разнородности классов клиентов и эффективности рекламной кампании?

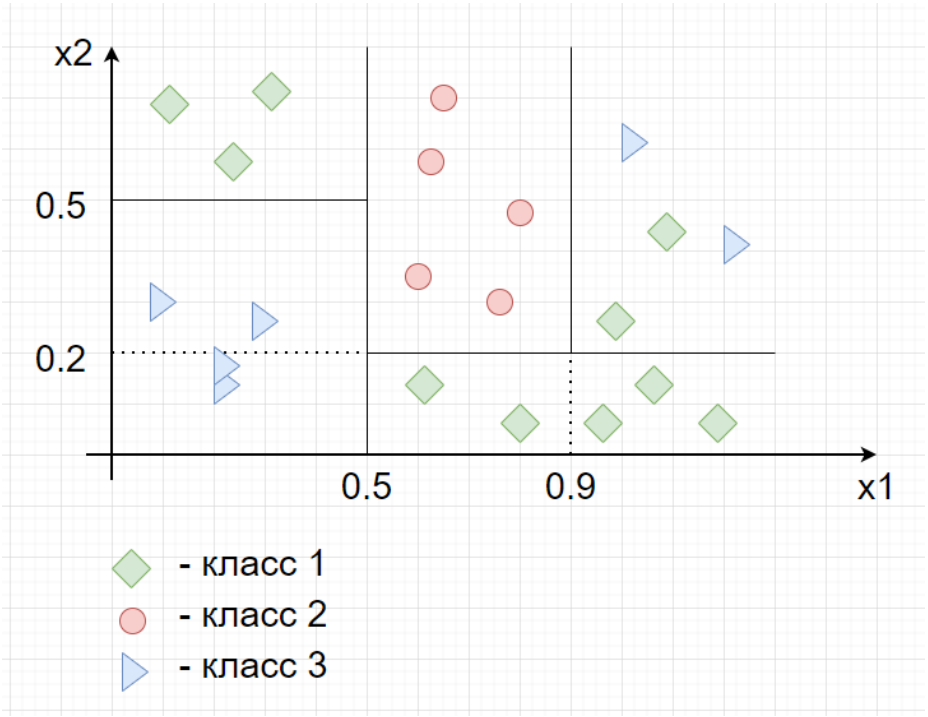
Правильный ответ:
Индекс Джини измеряет разнородность классов в наборе данных. Значение близкое к нулю указывает на более однородные классы, а значение близкое к единице - на более разнородные. Индекс Джини может быть использован для сравнения распределения классов до и после рекламной кампании. Уменьшение индекса может свидетельствовать о том, что кампания привлекла более разнообразные группы клиентов.

Критерии оценки:

Зачтено: Сформулирован вариант интерпретации индекса Джини применительно к рассматриваемой задаче (оценка эффективности рекламной кампании).

Не зачтено: Предложенный ответ не относится к рассматриваемой задаче или неправильно интерпретирует индекс Джини.

3. Для некоторого набора данных решается задача классификации методом деревьев решений. На рисунке представлено разбиение признакового пространства. Напишите правило в виде последовательности условий, для листа дерева, выделяющего 2 (второй) класс (на рисунке обозначен кружком) объектов



Правильный ответ:
x1>0.5 И x1<0.9 И x2>0.2

Критерии оценки:

Зачтено: Верно записана последовательность условий в виде предиката, указанного в правильном ответе или эквивалентного ему, либо перечислены условия через запятую или иной разделитель. Указание нестрого неравенства вместо строго не является ошибкой.

Не зачтено: записанный ответ не позволяет отделить указанный в задаче класс.

Тестовые задания закрытого типа (с одним вариантом ответа):

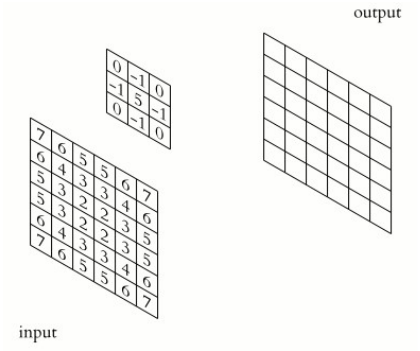
1. К основным характеристикам Big Data относятся:
а. Virtualization, Volume, Variability, Vehicle
б. Variety, Velocity, Volume, Value
в. Verification, Volume, Velocity, Visualization
г. Video, Value, Variety, Volume

2. **НЕВЕРНО** утверждение, что MapReduce:
а. интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
б. две операции: распределения и сборки данных
в. придуман разработчиками Hadoop
г. анонсирован разработчиками Google

3. Hadoop – это:
а. набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах
б. распределённая СУБД, позволяющая обрабатывать большие данные
в. язык выполнения заданий в парадигме MapReduce
г. распределённая файловая система, предназначенная для хранения файлов большого объёма

4. Задача поиска групп объектов в не размеченном исходном наборе данных - это:
а. классификация
б. кластеризация
в. регрессия
г. ассоциация

5. На рисунке



представлена нейронная сеть:

а. прямого распространения
б. сверточная
в. рекуррентная
г. обратного распространения

6. К классу «обучение с учителем» относят задачи интеллектуального анализа данных:
Возможно несколько вариантов ответа
а. классификация
б. кластеризация
в. регрессия
г. поиск ассоциативных правил

7. В процессе управляемого обучения нейронной сети минимизации требует целевая функция:
а. ошибок
б. переобучения
в. активации
г. корреляции

Тестовые задания закрытого типа (на последовательность, соответствие и с множеством вариантов правильных ответов):

1. Укажите в правильном порядке этапы методологии CRISP-DM.
а. Моделирование

б. Оценка соответствия целям
в. Внедрение
г. Понимание бизнес-целей
д. Понимание данных
е. Подготовка данных
Ответ: г, д, е, а, б, в.

2. Подготовка данных (Data Preparation) включает
а. консолидацию данных;
б. агрегирование и формирование выборок;
в. обогащение и очистку данных;
г. выбор модели данных
Ответ: а, б, в.

3. Преимуществами алгоритма CART (Classification and Regression Tree) являются
а. не требует вычисления параметров вероятностных распределений;
б. атрибуты разбиения выбираются непосредственно в процессе построения дерева,
в. изменения в обучающем множестве порождают значительные изменения в структуре дерева решений;
г. устойчив к выбросам и аномальным значениям.
Ответ: а, б, г.

4. Деревья решений, как класс алгоритмов классификации, имеют следующие недостатки:
а. неустойчивость, т.к. изменения в обучающем множестве ведет к изменению структуры дерева
б. могут работать только с категориальными данными
в. алгоритмы являются жадными и не дают гарантированного глобально оптимального решения
г. быстрое переобучение
Ответ: а, в, г.

5. Функция D(x,y) называется метрикой (расстоянием), если выполняются следующие условия
а. $D(x,y)>=0$
б. $D(x,y)=D(y,x)$
в. $D(x,z)<=D(x,y)+D(y,z)$
г. $D(x,y)=D(x,x)+D(y,y)$
Ответ: а, б, в.

6. Укажите в правильном порядке этапы первой итерации алгоритма k-means
а. Для каждого наблюдения исходного множества определяется ближайший к нему центр кластера
б. Вычисляются центроиды — центры тяжести кластеров. Каждый центроид — это вектор, элементы которого представляют собой средние значения соответствующих признаков, вычисленные по всем записям кластера.
в. Центр кластера смещается в его центроид, после чего центроид становится центром нового кластера.
г. Выбирается число кластеров k.
д. Из исходного множества данных случайным образом выбираются k наблюдений, которые будут служить начальными центрами кластеров.
Ответ: г, д, а, б, в.

4 ТИПОВЫЕ ЗАДАНИЯ НА КОНТРОЛЬНУЮ РАБОТУ, КУРСОВУЮ РАБОТУ/КУРСОВОЙ ПРОЕКТ

Данный вид контроля по дисциплине не предусмотрен учебным планом.

5 СВЕДЕНИЯ О ФОНДЕ ОЦЕНОЧНЫХ СРЕДСТВ И ЕГО СОГЛАСОВАНИИ

Фонд оценочных средств для аттестации по дисциплине «Технологии Data Mining» представляет собой компонент основной профессиональной образовательной программы магистратуры по направлению подготовки 09.04.01 Информатика и вычислительная техника.

Фонд оценочных средств рассмотрен и одобрен на заседании методической комиссии института цифровых технологий (протокол № 2 от 26.04.2022 г.).

Фонд оценочных средств актуализирован. Изменения, дополнения рассмотрены и одобрены на заседании методической комиссии института цифровых технологий (протокол № 3 от 24.03.2023 г.).

Директор института  А.Б. Тристанов