


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БАЙКАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ЧИТИНСКИЙ ИНСТИТУТ
КОЛЛЕДЖ

УТВЕРЖДАЮ:
Первый заместитель директора

Н.В. Раевский
«25» июня 2024 г.

**ФОНД
ОЦЕНОЧНЫХ СРЕДСТВ**

по дисциплине
ОП.17 Анализ и обработка информации

Специальность 09.02.07 Информационные системы и программирование

Квалификация: Программист

Чита 2024 г.

**Структура
фонда оценочных средств
по дисциплине «Анализ и обработка информации»**

№ п/п	Тема из рабочей программы	Закрываемая компетенция	Формируемые ЗУ-Ны (З.1, У1)	Вид и номер задания в ФОС	Критерий оценивания (по 100-% шкале)
1.	Раздел 1. Введение	ОК 2.	У.1. определять задачи для поиска информации	Лабораторная работа №1	15
	Раздел 2. Основные понятия, используемые при анализе информации				
2.	2.1. Большие данные и машинное обучение	ОК 2. ПК 2.5.	У.2. определять необходимые источники информации; У.4. структурировать получаемую информацию; У.5. выделять наиболее значимое в перечне информации; У.6. оценивать практическую значимость результатов поиска; У.7. оформлять результаты поиска. 3.1. приемы структурирования информации; 3.2. формат оформления результатов поиска информации; 3.4. возможности алгоритмов машинного обучения; 3.5. классы задач, решаемых с помощью алгоритмов машинного обучения.	Тест по теме (20 вопросов)	17-20 верных ответов – «5» 11-16 верных ответов – «4» 5-10 верных ответов – «3» 4 и менее – «2»
3.	2.2. Введение в программирование на языке Python	ОК 2. ПК 2.5.	У.3. планировать процесс поиска; У.8. проводить анализ данных;	Тест по теме (10 вопросов)	10-9 верных ответов – «5» 8-7 верных ответов – «4» 6-5 верных ответов – «3» 4 и менее – «2»
4.	2.3 Метрические методы классификации	ОК 2. ПК 2.5.	У.6. оценивать практическую значимость результатов поиска; У.9. применять на практике алгорит-	Лабораторная работа №2	15

			<p>мы машинного обучения; У.10. обосновать применение того или иного алгоритма машинного обучения для решения конкретной задачи. 3.3. методы интеллектуального анализа данных (включая их преобразование и очистку, работу с пропущенными значениями, основные способы визуализации данных, корреляционный анализ, поиск нелинейных ассоциаций);</p>		
5.	2.4. Логические методы классификации	ОК 2. ПК 2.5.	<p>У.6. оценивать практическую значимость результатов поиска; У.9. применять на практике алгоритмы машинного обучения; У.10. обосновать применение того или иного алгоритма машинного обучения для решения конкретной задачи. 3.3. методы интеллектуального анализа данных (включая их преобразование и очистку, работу с пропущенными значениями, основные способы визуализации данных, корреляционный анализ, поиск нелинейных ассоциаций);</p>	Лабораторная работа №3	15
6.	2.5. Линейные методы классификации	ОК 2. ПК 2.5.	<p>У.6. оценивать практическую значимость результатов поиска; У.9. применять на</p>	Лабораторная работа №4	15

			<p>практике алгоритмы машинного обучения;</p> <p>У.10. обосновать применение того или иного алгоритма машинного обучения для решения конкретной задачи.</p> <p>З.3. методы интеллектуального анализа данных (включая их преобразование и очистку, работу с пропущенными значениями, основные способы визуализации данных, корреляционный анализ, поиск нелинейных ассоциаций);</p>		
12	Итого по текущей аттестации	ОК 2. ПК 2.5.	<p>Сформированы: 3.1; 3.2; 3.3; 3.4; 3.5 У.1; У.2; У.3; У.4; У.5; У.6; У.7; У.8; У.9; У.10</p>	<p>4 – лабораторных работ; 15 – тестовых заданий.</p>	<p><i>Количество баллов минимальное 40; максимальное 70</i></p>
13.	Промежуточная аттестация	ОК 2. ПК 2.5.	<p>У.4. структурировать получаемую информацию;</p> <p>У.5. выделять наиболее значимое в перечне информации;</p> <p>У.6. оценивать практическую значимость результатов поиска;</p> <p>У.8. проводить анализ данных;</p> <p>З.3. методы интеллектуального анализа данных (включая их преобразование и очистку, работу с пропущенными значениями, основные способы визуализации данных, корреляционный анализ, поиск нелинейных ассоциаций);</p>	Итоговый тест 30 вопросов	<p>Зачет проводится в форме компьютерного тестирования. Каждый правильный ответ на вопрос теста оценивается в 1 балл</p> <p>«5»- 24-30 ответов;</p> <p>«4» -17-23 ответов</p> <p>«3»- 10-16 ответов</p>

			3.4. возможности алгоритмов машинного обучения; 3.5. классы задач, решаемых с помощью алгоритмов машинного обучения.		
--	--	--	---	--	--

Результаты освоения дисциплины, подлежащие проверке

Обучающийся в ходе освоения дисциплины должен:

уметь:

- У.1. определять задачи для поиска информации;
- У.2. определять необходимые источники информации;
- У.3. планировать процесс поиска;
- У.4. структурировать получаемую информацию;
- У.5. выделять наиболее значимое в перечне информации;
- У.6. оценивать практическую значимость результатов поиска;
- У.7. оформлять результаты поиска.
- У.8. проводить анализ данных;
- У.9. применять на практике алгоритмы машинного обучения;
- У.10. обосновать применение того или иного алгоритма машинного обучения для решения конкретной задачи.

знать:

- З.1. приемы структурирования информации;
- З.2. формат оформления результатов поиска информации;
- З.3. методы интеллектуального анализа данных (включая их преобразование и очистку, работу с пропущенными значениями, основные способы визуализации данных, корреляционный анализ, поиск нелинейных ассоциаций);
- З.4. возможности алгоритмов машинного обучения;
- З.5. классы задач, решаемых с помощью алгоритмов машинного обучения.

Изучение дисциплины способствует освоению **общих и профессиональных компетенций:**

Код	Наименование результатов обучения
ОК 02	Осуществлять поиск, анализ и интерпретацию информации, необходимой для выполнения задач профессиональной деятельности.
ПК 2.5	Производить инспектирование компонентов программного обеспечения на предмет соответствия стандартам кодирования

Комплект заданий для лабораторных работ
по дисциплине «Анализ и обработка информации»

Основы работы с пакетом KNIME

KNIME представляет собой свободно распространяемый прикладной программный пакет с графическим интерфейсом, поддерживающий цикл интеллектуального анализа данных (доступ к данным различных форматов, трансформация данных, аналитические функции, визуализация и подготовка отчетов).

Идеологической основой KNIME является понятие потока работ (workflow). Поток работ графически изображает процесс преобразования исходных данных в результаты (см. Рис. 1). Изображение состоит из узлов (прямоугольников) и стрелок. Узел инкапсулирует некоторую операцию над данными, стрелки показывают путь данных.

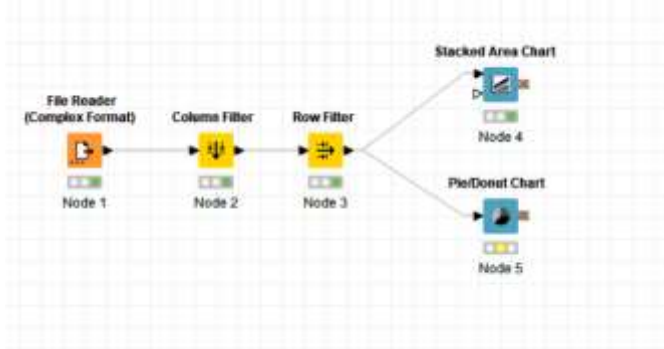
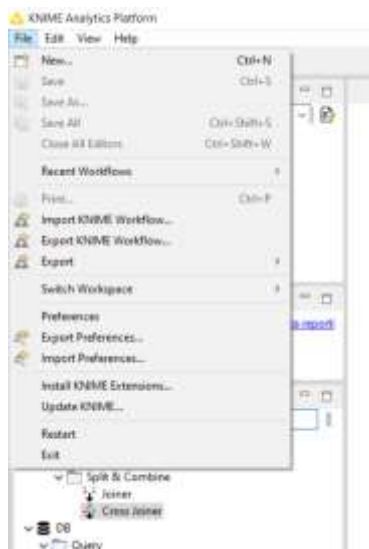


Рис. 1. Простой поток работ

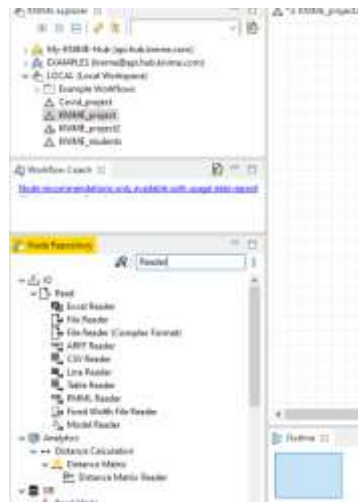
Скачайте и установите KNIME Analytics Platform

1. Для начала скачайте sales_data.csv, содержащий данные, которые вы будете использовать в рабочем процессе. Откройте платформу KNIME и создайте новый пустой рабочий процесс, нажав «New...» на панели инструментов и далее «New KNIME Workflow».

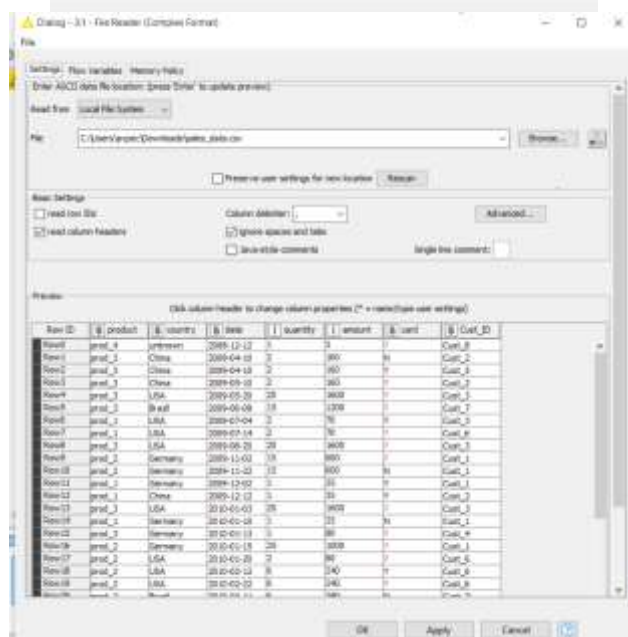
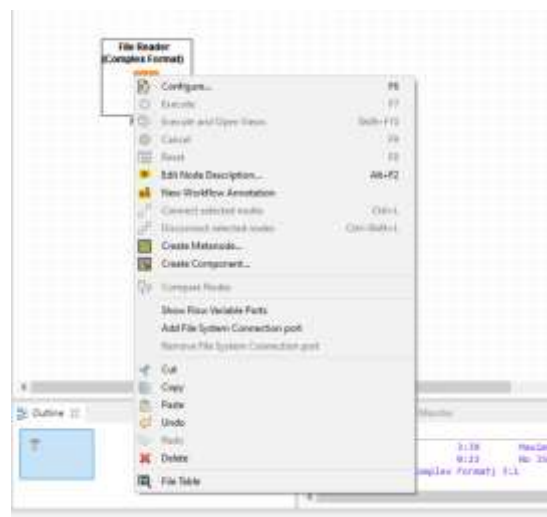


Назовите проект и нажмите кнопку «Finish».

2. В поиске окна Node Repository введите слово «Reader» (посмотрите какие типы файлов может считывать программа) перетащите элемент File Reader (Complex Format) в редактор рабочей среды.



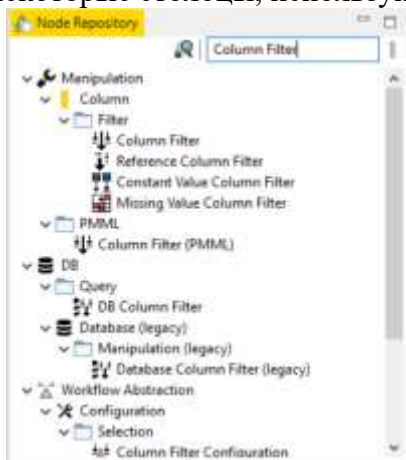
Откройте контекстное меню и нажмите «Configure...»



Здесь Вы можете выбрать путь к файлу, который хотите использовать для обработки, после этого Вы увидите предварительный просмотр таблицы данных. Нажмите Apply и OK и запустите узел File Reader (Complex Format), щелкнув узел правой кнопкой мыши

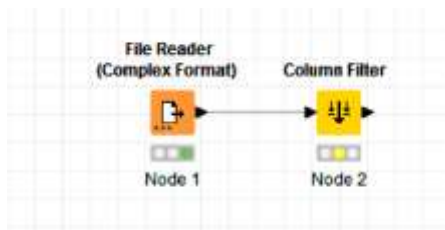
и выбрав Execute в контекстном меню. Теперь входные данные доступны на выходном порту узла File Reader (Complex Format). Чтобы просмотреть выходную таблицу, щелкните правой кнопкой мыши исполняемый узел и выберите в меню последний пункт: File Table.

3. Чтобы отфильтровать некоторые столбцы, используйте узел «Column Filter».



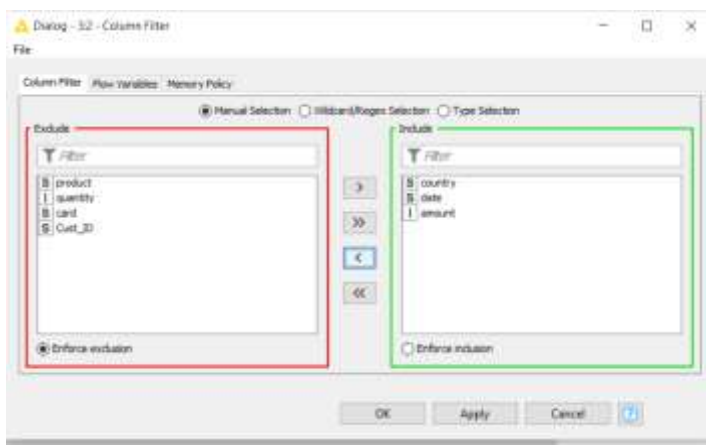
На панели репозитория узлов слева:

- напишите «Column Filter» в поле поиска
- перетащите Column Filter в редактор рабочего процесса.
- соедините его входной порт с выходным портом узла File Reader (Complex Format)



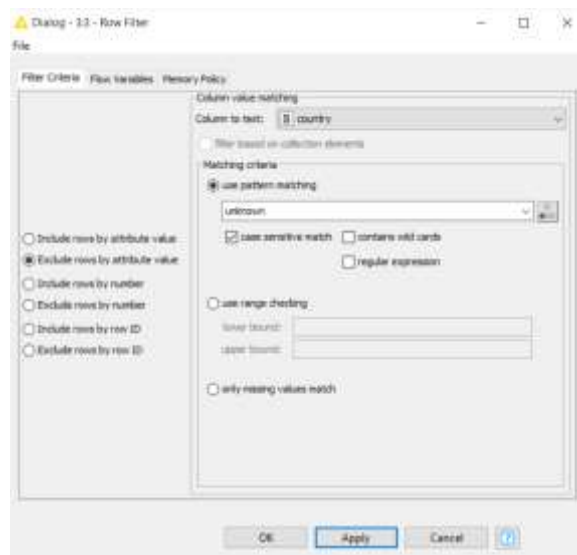
Чтобы открыть диалоговое окно конфигурации, щелкните узел правой кнопкой мыши и выберите «Configure...».

Здесь переместите столбцы: «страну», «дату» и «сумму» в поле «Include» в правой части диалогового окна, затем нажмите «ОК». После выполнения узла таблица отфильтрованных данных доступна на выходном порту узла Column Filter.



4. Чтобы очистить ваши данные, отфильтровав строки, соответствующие определенным значениям одного столбца, используйте узел Row Filter. Найдите узел Row Filter в репозитории узлов слева, добавьте его в рабочий процесс и соедините с узлом Column Filter.

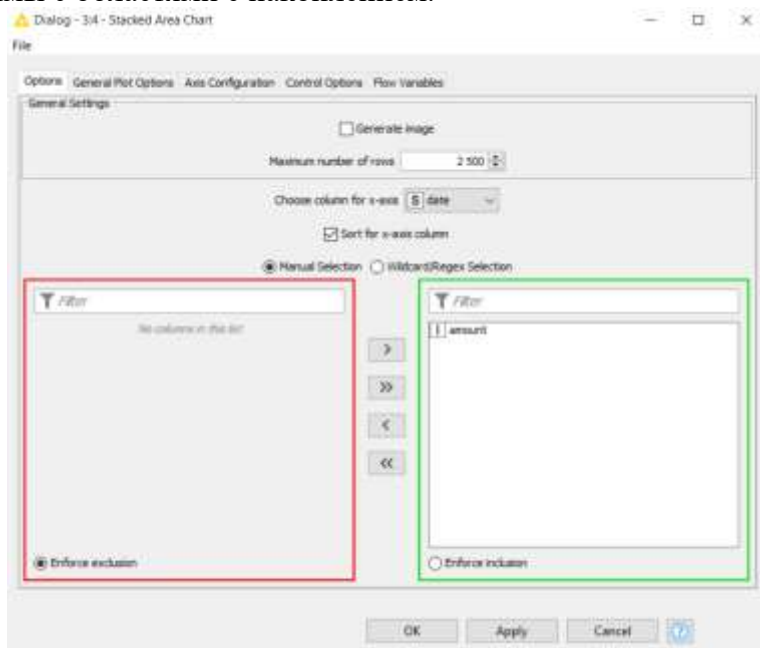
Откройте диалоговое окно конфигурации узла Row Filter и исключите строки из входной таблицы, где столбец «страна» имеет значение «неизвестно».



Нажмите ОК и выполните узел. Теперь таблица отфильтрованных данных доступна на выходном порту узла Row Filter.

5. Чтобы визуализировать данные, например, построить диаграмму с областями с накоплением и круговую диаграмму, используйте узлы Stacked Area Chart (Диаграмма с областями с накоплением) и Pie/Donut Chart (Круговая/кольцевая диаграмма). Найдите их и добавьте в рабочий процесс, подключив оба к узлу Row Filter.

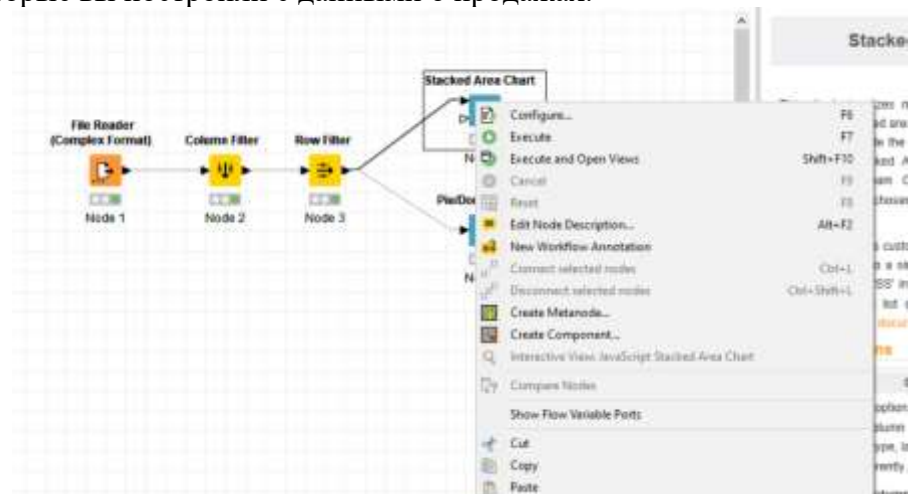
Настройте столбец «дата» как столбец оси X в диалоговом окне конфигурации узла диаграммы с областями с накоплением.



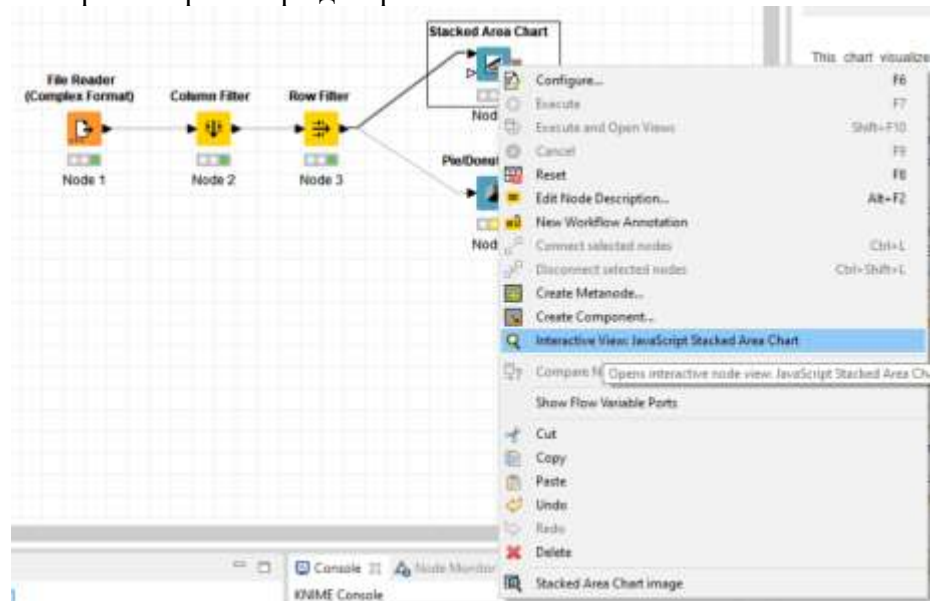
Настройте столбец «страна» в качестве столбца категории, «сумма» в качестве метода агрегирования и «количество» в качестве столбца частоты в узле круговой/кольцевой диаграммы.



Теперь вы можете выполнить и просмотреть вывод этих двух последних узлов. Щелкните узел правой кнопкой мыши и выберите «Выполнить и открыть представление» в контекстном меню. Откроется новое окно, показывающее диаграммы, которые вы построили с данными о продажах.



Для повторного просмотра диаграмм нажмите Interactive View



Простой поток работ готов. Теперь можете ознакомиться с основными документами и перейти к лабораторным работам.

Изучите краткое Руководство пользователя KNIME. Проверьте себя на понимание основных терминов KNIME: поток работ (workflow), рабочее пространство (workspace), узел (node), порт узла (port), статус узла (node status), соединение и настройка узла (connecting and configuring node).

Лабораторная работа № 1. Поиск шаблонов

Цель. Построение потока работ, выполняющего решение задачи анализа рыночной корзины и поиска ассоциативных правил. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, преобразовать загруженные данные в специализированный тип данных пакета KNIME, найти частые наборы и ассоциативные правила, вывести результаты.

1. Создайте поток работ, приведённый на Рис. 2.

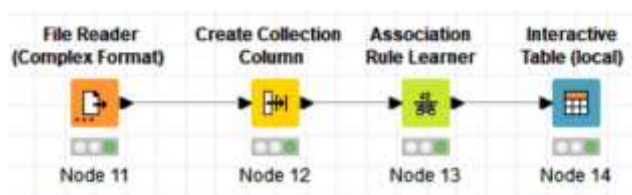


Рис. 2. Поток работ для решения задачи анализа рыночной корзины

2. Выполните настройку узла “File Reader” так, чтобы:

- в качестве файла исходных данных фигурировал baskets.csv;
- первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”);
- в качестве разделителя столбцов фигурировала запятая;
- узел обрабатывал неполные строки (кнопка “Advanced”, вкладка “Short lines”).

3. Выполните настройку узла “Create Collection Column” так, чтобы:

- все строки исходного файла попали в выходную коллекцию данных;
- установите флаги “Create collection of type set”, “Ignore missing values”, “Remove aggregated columns from table”.

4. Выполните поток работ, предварительно настроив узел “Association Rule Learner”, указав различные значения поддержки (minimum support). Объясните полученные результаты (как данный параметр влияет на решение задачи?).

5. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Лабораторная работа № 2. Деревья решений

Цель. Построение потока работ, выполняющего решение задачи классификации посредством построения дерева решений. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

1. Создайте поток работ, приведённый на Рис. 2.

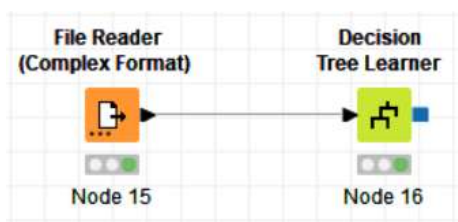


Рис. 3. Поток работ для решения задачи классификации

2. Выполните настройку узла “File Reader” так, чтобы:

- в качестве файла исходных данных фигурировал marks.csv;
- первая строка файла трактовалась как содержащая названия столбцов (“Read column headers”);

3. Выполните настройку узла “Decision Tree Learner” так, чтобы поле FINALMARK трактовалось как признак класса.

4. Выполните поток работ. Убедитесь, что построенное дерево решений показывает зависимость итоговой оценки только от атрибутов, отражающих мнение учителя (имеющих название вида TEACHER_xx), и ее независимость от атрибутов, отражающих персональные данные ученика (имеющих название вида PUPIL_xx).

5. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

6. Измените созданный поток работ, как показано на Рис. 4.

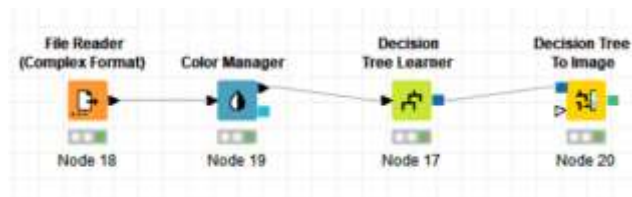


Рис. 4. Поток работ для решения задачи классификации с дополнительными узлами

1. Настройте узел “Color Manager”, указав столбец FINALMARK как раскрашенный. Выполните поток работ. Сравните результаты при различных настройках узла “Decision Tree Learner” (контекстное меню узла): “Decision Tree View” и “Decision Tree View (simple)”.

2. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Лабораторная работа № 3. Кластеризация

Цель. Построение потока работ, выполняющего решение задачи кластеризации. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

1. Создайте поток работ, приведённый на Рис. 5.

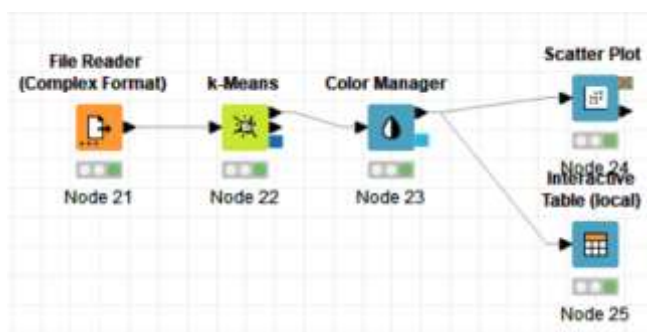


Рис. 5. Поток работ для решения задачи кластеризации

2. Выполните настройку узла “File Reader” так, чтобы в качестве файла исходных данных фигурировал basketball.csv.

3. Выполните поток работ. Сравните результаты при различных параметрах: количество кластеров, цвета для отображения кластеризуемых объектов.

4. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Лабораторная работа № 4. Построение рабочего процесса предварительной обработки данных

Целью этого упражнения является создание простого рабочего процесса для предварительной обработки данных. Мы будем использовать анонимизированные данные переписи населения США 1994 г., где каждый кортеж данных состоит из атрибутов, перечисленных ниже.

1.	AGE	Возраст личности.
2.	WORKCLASS	Тип занятости (возможные значения Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Localgov, State-gov, Without-pay, Never-worked).
3.	FNLWGT	вес, присвоенный Бюро переписи населения.
4.	EDUCATION	Высший уровень образования, достигнутый этим человеком (возможные значения Bachelor, Some-college, Master, Doctorate, etc.).
5.	EDUCATION-NUM	Высший уровень образования в числовой форме.
6.	MARITAL-STATUS	Семейное положение физического лица (возможные значения Marriedciv-spouse, Divorced, Never-married, etc.).
7.	OCCUPATION	Профессия человека (возможные значения Techsupport, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Farming-fishing, Transportmoving, Armed-Forces, etc.).
8.	RELATIONSHIP	Семейные отношения личности (возможные значения Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).
9.	RACE	Раса человека (возможные значения White, Asian-PacIslander, Amer-Indian-Eskimo, Black, Other).
10.	SEX	Пол человека (возможные значения Female, Male).
11.	CAPITAL-GAIN	прирост капитала физического лица.
12.	CAPITAL-LOSS	потери капитала физического лица.
13.	HOURS-PER-WEEK	Количество часов, отработанных человеком в неделю.
14.	NATIVE-COUNTRY	Страна происхождения для физического лица (возможные значения UnitedStates, Cambodia, England, Puerto-Rico, Canada, Germany, etc.).
15.	INCOME	Флаг, показывающий, зарабатывает ли человек более 50000 долларов в год. (возможные значения >50K, <=50K).

Вы должны последовательно выполнить следующие шаги предварительной обработки:

- Разделить входную таблицу по вертикали на две таблицы: Первая, с именными столбцами только (т.е. WORKCLASS, EDUCATION, MARITAL-STATUS, OCCUPATION, RELATIONSHIP, RACE, SEX, INCOME) и второе, только с числовыми столбцами (например, AGE, FNLWGT, EDUCATION-NUM, CAPITAL-GAIN, CAPITAL-LOSS, HOURS-PER-WEEK).

- Сохранить все строки первой таблицы, где EDUCATION — Bachelor или Master, и удалить другие.

- Сохранить все строки второй таблицы, где AGE от 21 до 65 лет, и удалить другие.

- Присоединяйтесь к первой и второй таблицам, чтобы обеспечить соблюдение обоих критериев, упомянутых выше.

- Рассчитать различную статистику (минимум, максимум, среднее и т. д.) результирующей таблицы из предыдущий шаг.
- Сохраните столбцы ВОЗРАСТ и ЧАСЫ В НЕДЕЛЮ второй таблицы и удалить другие.
- Нарисуйте блочную диаграмму для результирующей таблицы из предыдущего шага.

Выполнение

1. Создайте поток работ, приведённый на Рис. 56.

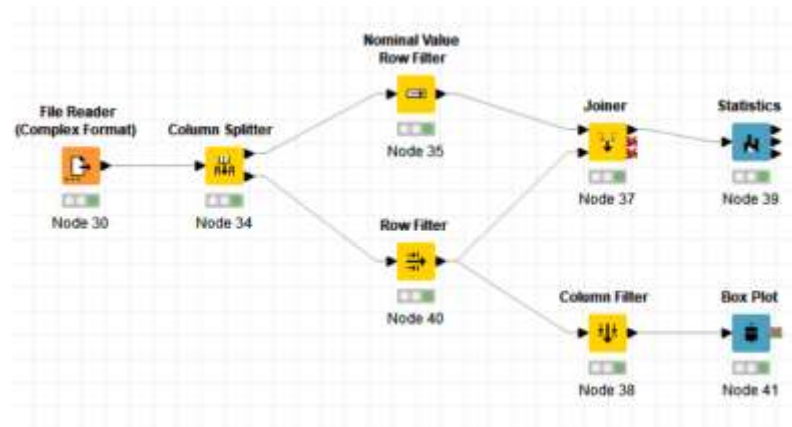


Рис. 6. Поток работ для решения задачи предварительной обработки данных

2. Выполните настройку узла “File Reader” так, чтобы в качестве файла исходных данных фигурировал adult.csv.
3. Настройте узлы по заданию, если появляются вопросы – обратитесь к преподавателю.
4. Выполните поток работ. Сравните результаты при различных параметрах: количество кластеров, цвета для отображения кластеризуемых объектов.
5. Создайте скриншоты потока работ и результатов его работы для использования в качестве отчета о выполнении задания.

Комплект тестов

ТЕСТ

по теме «Большие данные и машинное обучение»

1. К какому типу задач машинного обучения относится задача предсказания цены жилья по его характеристикам?
 - a) Классификация на два класса;
 - b) Классификация на M непересекающихся классов;
 - c) Классификация на M пересекающихся классов;
 - d) Восстановление регрессии;
2. К какому типу относится признак «Цвет глаз»?
 - a) Бинарный;
 - b) Количественный;
 - c) Номинальный (категориальный);
 - d) Порядковый;
3. К какому типу задач машинного обучения, относится задача в которой необходимо определить независимые группы и их характеристики во всем множестве анализируемых данных?
 - a) задача классификации;
 - b) задача регрессии;
 - c) задача кластеризации;
4. К какому типу задач машинного обучения, относится задача в которой необходимо определить зависимости между объектами или событиями?
 - a) задача распознавания образов;
 - b) задача поиска ассоциативных правил;
 - c) задача нормализации;
5. Астроном Витя хочет построить модель, которая сможет разбить известные науке звезды на группы по их характеристикам, чтобы лучше изучить их особенности. К какому типу относится данная задача?
 - a) Кластеризация
 - b) Классификация
 - c) Ранжирование
 - d) Регрессия
6. Выберите все верные утверждения:
 - a) Модель машинного обучения, по сути, является отображением пространства ответов в пространство объектов
 - b) Функционал ошибки показывает, насколько плохое качество имеют данные, используемые для решения задачи
 - c) Процесс обучения модели заключается в минимизации функционала ошибки

d) Элементами обучающей выборки являются объекты, характеристики которых являются значениями признаков

7. Можно ли из модели линейной регрессии выбросить свободный коэффициент ω_0 и почему?

- a) Нет, потому что без ω_0 модель будет константной
- b) Нет, потому что без свободного коэффициента модель гарантированно будет давать нулевой прогноз при нулевых значениях всех признаков, а это ограничивает возможности по подгонке под данные
- c) Да, он участвует в модели только для упрощения процедуры обучения, но при этом никак не повышает её силу
- d) Да, он участвует в модели исключительно по историческим причинам

8. Рассмотрим признак “Образовательная программа” при анализе данных по студентам университета. Этот признак может принимать три значения: “Экономика”, “Математика”, “Философия”. Воспользуемся one-hot кодированием и заменим этот признак на три бинарных, которые будут соответствовать категориям в том порядке, в котором они перечислены выше. Как будет закодирован признак со значением “Философия”?

- a) (0, 1, 0)
- b) (1, 0, 0)
- c) (0, 0, 1)

9. Предположим, что мы строим модель предсказания роста по возрасту и весу человека.

Модель с какими коэффициентами вероятнее всего переобучилась?

- a) $0.001 * (\text{возраст}) + 0.5 * (\text{вес})$
- b) $1402325.3 * (\text{возраст}) + -1404370.5 (\text{вес})$ верно
- c) $0.1 * (\text{возраст}) + 0.33 (\text{вес})$

10. Предположим, что мы строим модель предсказания стоимости дома по количеству комнат и средней цене дома в районе. Перед количеством комнат коэффициент равен 1400230, а перед средней ценой дома в районе 0.8. Можно ли утверждать, что количество комнат — более важный признак для качества предсказания, чем средняя цена в районе и почему?

- a) Да, так как количество комнат — это признак, который может принимать небольшое количество значений, а значит, каждое значение содержит в себе больше информации
- b) Нет, так как коэффициенты несравнимы, поскольку признаки имеют разный масштаб
- c) Нет, так как средняя цена дома в районе — это признак с большим разбросом, а именно разброс характеризует ценность признака
- d) Да, так как коэффициент перед количеством комнат больше

11. Заполните пропуск: «Для выпуклой функции указывает сторону наискорейшего убывания»

- a) Изохрона
- b) Линия уровня
- c) Антиградиент
- d) Градиент

12. Позволяет ли градиентный спуск находить минимум функции?

- a) Нет, процесс подбора не сходится к минимуму

- b) Да, алгоритм всегда сходится к глобальному минимуму
- c) Да, но только локальный минимум

13. Найдите евклидову норму вектора $x = (1, 1, 1.5, 1)$. Ответ вводите с точкой, с точностью до 2-х знаков после запятой.

14. В лекции был рассказан алгоритм градиентного спуска, в котором на каждом шаге параметры сдвигаются в сторону антиградиента, за счёт чего уменьшается ошибка. Представьте себе другой алгоритм: на каждом шаге мы выбираем один из параметров и сдвигаемся по нему влево или вправо (выбираем такое направление, чтобы ошибка уменьшалась сильнее). Такой метод называют покоординатным спуском. Чем он хуже?

- a) Покоординатный спуск, скорее всего, потребует большего числа итераций, чем градиентный спуск.
- b) Покоординатный спуск требует подсчёта градиентов, а это может быть достаточно затратно по времени.
- c) Покоординатный спуск не даёт никаких гарантий того, что ошибка будет уменьшаться с числом шагов.
- d) Покоординатный спуск может быть неэффективным в пространствах большой размерности, поскольку даже чтобы сдвинуться по каждой координате по одному разу, потребуется много шагов.

15. Чем глобальный минимум отличается от локального?

- a) В точке глобального минимума функция принимает уникальное значение, которое не достигается больше нигде.
- b) В точке глобального минимума функция принимает самое маленькое значение по сравнению с точками, находящимися рядом.
- c) В точке глобального минимума функция принимает такое значение, что меньше него достичь нельзя.

16. Чем стохастический градиентный спуск (SGD) лучше обычного градиентного спуска?

Выберите все подходящие ответы.

- a) Один шаг в SGD быстрее, поэтому в целом этот метод может быстрее выдать решение.
- b) В SGD гарантируется, что на каждой итерации уменьшается ошибка модели.
- c) В SGD гарантируется, что будет найден глобальный минимум.
- d) Один шаг в SGD точнее, чем в обычном градиентном спуске, поэтому требуется меньше шагов для получения решения.

17. Вы решаете задачу классификации писем на нормальные и спам, причём примеров спама очень мало (меньше 1% от размера всей выборки). Вы решили рассмотреть модель, которая для любого письма говорит, что оно нормальное. Выберите все верные утверждения про метрики качества для такой модели.

- a) Доля верных ответов будет близка к 50%
- b) Полнота будет равна единице
- c) Доля верных ответов будет высокой
- d) Точность будет равна единице

18. Что оценивает абсолютное значение отступа?

- a) Модуль веса, соответствующего данному объекту
- b) Вероятность класса для данного объекта
- c) Расстояние от объекта до разделяющей гиперплоскости

- d) Вероятность того, что данный объект является выбросом

19. Рассмотрим выборку из трех объектов, принадлежащих к классам 0, 1 и 1 соответственно. Оценка принадлежности классу 1 алгоритма классификации для первого объекта равна 0.2, для второго - 0.4, и для третьего - 0.9. Найдите площадь под PR-кривой для данного классификатора на данной выборке.

20. Выберите верные утверждения про метод опорных векторов для линейно неразделимого случая:

- a) В методе опорных векторов максимизируется расстояние от разделяющей поверхности до ближайшего объекта обучающей выборки
- b) В методе опорных векторов минимизируется расстояние от разделяющей поверхности до ближайшего объекта обучающей выборки
- c) В методе опорных векторов на объекте может быть допущена ошибка, но за это в функционале добавляется штраф
- d) В методе опорных векторов запрещено допускать ошибки на объектах обучающей выборки

ТЕСТ

по теме «Введение в программирование на языке Python»

1. Какие существуют типы переменных (выбрать несколько вариантов):
 - a) float
 - b) str
 - c) num
 - d) int
 - e) bool
 - f) real
2. Переменная int:
 - a) вещественная переменная
 - b) символьная строка
 - c) логическая переменная
 - d) целая переменная
3. Переменная float:
 - a) вещественная переменная
 - b) символьная строка
 - c) логическая переменная
 - d) целая переменная
4. Переменная str:
 - a) вещественная переменная
 - b) символьная строка
 - c) логическая переменная
 - d) целая переменная
5. Переменная bool:
 - a) вещественная переменная
 - b) символьная строка
 - c) логическая переменная
 - d) целая переменная

6. Имена переменных не могут включать:

- a) Русские буквы
- b) Латинские буквы
- c) Пробелы
- d) Скобки, знаки + = ! ? b др.
- e) Цифры

7. Что будет в результате выполнения программы:

```
a = 20
b = a + 4
a = b * 100
print(a)
```

- a) 240
- b) 2400
- c) 100
- d) 420

8. Что будет в результате выполнения следующего алгоритма:

Входные данные: 57

```
x = int(input())
if x > 0:
    print(x)
else:
    print(-x)
```

0
-57
57
23

9. Что будет в результате выполнения следующего алгоритма:

Входные данные: -57

```
x = int(input())
if x > 0:
    print(x)
else:
    print(-x)
```

- a) 0
- b) -57
- c) 57
- d) 23

10. Что будет в результате выполнения следующего алгоритма программы:
Входные данные:

15
45

```
a = int(input())  
b = int(input())  
if a % 10 == 0 or b % 10 == 0:  
    print('YES')  
else:  
    print('NO')
```

- a) YES
- b) NO

Итоговый тест

1. Задача классификации сводится к ...
 - a) нахождению частых зависимостей между объектами или событиями;
 - b) определению класса объекта по его характеристиками;
 - c) определению по известным характеристиками объекта значение некоторого его параметра;
 - d) поиска независимых групп и их характеристик во всем множестве анализируемых данных.
2. Задача регрессии сводится к ...
 - a) нахождению частых зависимостей между объектами или событиями;
 - b) определения класса объекта по его характеристиками;
 - c) определение по известным характеристиками объекта значение некоторого его параметра;
 - d) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
3. Задача кластеризации заключается в ...
 - a) нахождению частых зависимостей между объектами или событиями;
 - b) определения класса объекта по его характеристиками;
 - c) определение по известным характеристиками объекта значение некоторого его параметра;
 - d) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
4. Целью поиска ассоциативных правил является ...
 - a) нахождению частых зависимостей между объектами или событиями;
 - b) определения класса объекта по его характеристиками;
 - c) определение по известным характеристиками объекта значение некоторого его параметра;
 - d) поиска независимых групп и их характеристик в всем множестве анализируемых данных.
5. До предполагаемых моделей относятся такие модели данных:
 - a) модели классификации и последовательностей;
 - b) регрессивные, кластеризации, исключений, итоговые и ассоциации;
 - c) классификации, кластеризации, исключений, итоговые и ассоциации;
 - d) модели классификации, последовательностей и исключений.

6. Что такое визуализация данных?
- a) построение отчетов по имеющимся данным
 - b) представление информации в виде графиков, диаграмм, структурных схем и т. д.
 - c) вывод информации на экран компьютера
 - d) печать данных на твердом носителе
7. В описательных моделях относятся следующие модели данных:
- a) модели классификации и последовательностей;
 - b) регрессивные, кластеризации, исключений, итоговые и ассоциации;
 - c) классификации, кластеризации, исключений, итоговые и ассоциации;
 - d) модели классификации, последовательностей и исключений.
8. Модели классификации описывают ...
- a) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
 - b) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
 - c) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
 - d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
9. Когда необходимо сравнить значения нескольких наборов данных, графически изобразить отличия значения одних данных от других, показать изменения данных с течением времени, целесообразно создать
- a) Круговую диаграмму
 - b) Гистограмму
 - c) Линейчатую диаграмму
10. Модели последовательностей описывают ...
- a) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
 - b) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
 - c) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
 - d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
11. Регрессивные модели описывают ...
- a) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
 - b) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
 - c) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
 - d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
12. Виды лингвистической неопределенности:

- a) неточность измерений значений определенной величины, выполняемых физическими приборами;
- b) неопределенность значений слов (Многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (Синтаксическая и семантическая);
- c) случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость)
- d) неоднозначность смысла фраз (Синтаксическая и семантическая).

13. Модели исключений описывают ...

- a) исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;
- b) ограничения на данные анализируемого массива;
- c) закономерности между связанными событиями;
- d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.

14. Итоговые модели обнаружат ...

- a) исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;
- b) ограничения на данные анализируемого массива;
- c) закономерности между связанными событиями;
- d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.

15. Модели ассоциации проявляют ...

- a) исключительные ситуации в записях, которые резко отличаются произвольной признаку от основной множества записей;
- b) ограничения на данные анализируемого массива;
- c) закономерности между связанными событиями;
- d) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.

16. Виды физической неопределенности данных:

- a) неточность измерений значений определенной величины, выполняемых физическими приборами; случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью)
- b) неопределенность значений слов (Многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (Синтаксическая и семантическая);
- c) случайность (или наличие в внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость);
- d) неоднозначность смысла фраз (Синтаксическая и семантическая).

17. Очистка данных — ...

- a) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.
- b) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач
- c) объект, содержащий структурированные данные, которые могут оказаться полезными для развязку аналитического задачи

d) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему

18. Обогащение — ...

a) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

b) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач

c) объект, содержащий структурированные данные, которые могут оказаться полезными для развязку аналитического задачи

d) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

19. Консолидация — ...

a) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

b) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач

c) объект, содержащий структурированные данные, которые могут оказаться полезными для развязку аналитического задачи

d) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему

20. Транзакция — ...

a) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

b) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов

c) высокоуровневые средства отражения информационной модели и описания структуры данных

d) это установление зависимости дискретной выходной переменной от входных переменных

21. Метаданные — ...

a) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

b) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов

c) высокоуровневые средства отражения информационной модели и описания структуры данных

d) это установление зависимости дискретной выходной переменной от входных переменных

22. Классификация — ...

a) некоторый набор операций над базой данных, который рассматривается как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных

b) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов

c) высокоуровневые средства отражения информационной модели и описания структуры данных

d) это установление зависимости дискретной выходной переменной от входных переменных

23. Регрессия — ...

a) это установление зависимости непрерывной выходной переменной от входных переменных

b) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

c) выявление закономерностей между связанными событиями

d) это установление зависимости дискретной выходной переменной от входных переменных

24. Кластеризация — ...

a) это установление зависимости непрерывной выходной переменной от входных переменных

b) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

c) выявление закономерностей между связанными событиями

d) это установление зависимости дискретной выходной переменной от входных переменных.

25. Ассоциация — ...

a) это установление зависимости непрерывной выходной переменной от входных переменных

b) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов

c) выявление закономерностей между связанными событиями

d) это установление зависимости дискретной выходной переменной от входных переменных

26. Машинное обучение — ...

a) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных

b) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов

c) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.

d) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных

27. Аналитическая платформа — ...

- a) специализированный программный решение (или набор решений), который включает в себя все инструменты для извлечения закономерностей из сырых данных
- b) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- c) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.
- d) подразделение искусственного интеллекта изучающий методы построения алгоритмов, способных обучаться на данных

28. Обучающая выборка — ...

- a) эта группировка объектов (Наблюдений, событий) на основе данных, описывающих свойства объектов
- b) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат
- c) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, что и отвечает ему правильный выходной результат.
- d) выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности

29. Ошибка обучения — ...

- a) это ошибка, допущенная моделью на учебной множества.
- b) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестовой множества
- c) имена, типы, метки и назначения полей исходной выборки данных
- d) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат

30. Ошибка обобщения — ...

- a) это ошибка, допущенная моделью на учебной множества.
- b) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестовой множества
- c) имена, типы, метки и назначения полей исходной выборки данных
- d) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданный входной влияние, и соответствующий ему правильный выходной результат