

1. В числе ключевых признаков больших данных (Big Data) – ... (укажите 2 варианта ответа)
2. К Big Data относятся ... данные
3. Установите соответствие между типом данных и примером:
4. Технология хранения и обработки больших объемов разнородных данных, которая сохраняет данные в исходном виде без жесткой предварительной структуры, – это ...
5. Основное отличие Hadoop от традиционных систем обработки данных заключается в том, что Hadoop ...
6. Расположите в правильной последовательности этапы обработки больших данных:
7. Модель MapReduce используется для ...
8. ... данные – это данные, имеющие четко заданную фиксированную схему, обычно представленную в виде таблиц с полями и типами данных
9. Говоря о языках запросов SQL и NoSQL, можно утверждать, что ... (укажите 2 варианта ответа)
10. К вопросам безопасности больших данных (Big Data) относится задача ...
11. Компания анализирует потоковые данные с датчиков IoT в реальном времени. Данные поступают непрерывно, имеют большой объем и разный формат (числа, текст, сигналы). Какая технология наиболее подходит для обработки таких данных?
12. Говоря об истории и принципах организации среды R, можно утверждать, что язык программирования R ...
13. Говоря о различиях языков R и Python, можно утверждать, что ... (укажите 2 варианта ответа)
14. Установите соответствие между типом данных в R и характеристикой:
15. ... – это специальный тип данных в R для хранения категориальных переменных, представляющий собой вектор целых чисел с соответствующим набором уникальных меток, что оптимизирует память и статистический анализ
16. Основное отличие среды RStudio от базового интерфейса R заключается в том, что RStudio ...
17. Расположите в правильной последовательности этапы предварительной обработки данных в R:
18. Функция merge() в R используется для ...

19. ... в R – это двумерный объект, содержащий элементы одного типа, доступ к которым осуществляется по двум индексам (строка и столбец), предназначенный для выполнения линейно-алгебраических операций и статистических расчетов
20. Говоря об организации вычислений в R, можно утверждать, что ... (укажите 2 варианта ответа)
21. Говоря о векторах, таблицах и датафреймах в R, можно утверждать, что ...
22. Аналитик загрузил в среду R таблицу с данными о продажах (100 000 строк). Ему нужно удалить строки с пропусками, отсортировать таблицу по убыванию выручки и построить график зависимости выручки от времени. Приведите правильный порядок действий при выполнении этих задач.
23. В основе алгоритма Apriori для поиска ассоциативных правил лежит ...
24. Для оценки качества ассоциативных правил используются такие метрики, как ... (укажите 2 варианта ответа)
25. Установите соответствие между понятием и его определением:
26. Метод анализа, позволяющий выявлять часто встречающиеся наборы товаров в чеках, называется анализом потребительской ...
27. Основное отличие алгоритмов поиска последовательных шаблонов от алгоритмов поиска ассоциативных правил заключается в том, что первые ...
28. Расположите в правильной последовательности этапы поиска ассоциативных правил с помощью алгоритма Apriori:
29. Сфера, в которой ассоциативные правила наиболее часто применяются для анализа потребительской корзины, – это ...
30. ... – это мера качества ассоциативного правила, показывающая, насколько часто правило оказывается верным с учетом случайного совпадения, и вычисляемая как отношение ожидаемой частоты ошибок к фактической
31. Говоря об информационной базе в контексте применения метода ассоциативных правил, можно утверждать, что ... (укажите 2 варианта ответа)
32. Функция ... из пакета arules в языке R используется для поиска частых наборов элементов и ассоциативных правил
33. Сетевой магазин анализирует историю заказов своих клиентов. Менеджеры хотят выявить, какие товары часто покупают вместе, чтобы разместить их на одной витрине. Данные содержат 500 000 транзакций за последний год. Какая технология наиболее подходит для решения этой задачи?

34. К задачам классификации без обучения относится ...

35. К итеративным методам кластерного анализа, в частности, относится метод ... (укажите 2 варианта ответа)

36. Установите соответствие между методом и его описанием:

37. Английская аббревиатура ... обозначает метод главных компонент, используемый для снижения размерности признакового пространства, который строит новые некоррелированные признаки, максимизирующие дисперсию исходных данных

38. Основное отличие метода опорных векторов (SVM) от дискриминантного анализа заключается в том, что SVM:

39. Расположите в правильной последовательности этапы построения главных компонент (PCA) в R:

40. Функция ... из пакета cluster в R используется для реализации метода k-медоидов

41. Если имеется матрица исходных данных размерностью 1 000 строк и 50 столбцов, а после применения метода главных компонент (PCA) было оставлено 5 главных компонент, которые объясняют 85 % суммарной дисперсии, – тогда размерность признакового пространства уменьшилась в ... раз (с округлением до целых чисел)

42. Говоря о бинарных классификаторах, можно утверждать, что ... (укажите 2 варианта ответа)

43. Пакет ... в R предоставляет унифицированный интерфейс для построения и оценки различных моделей машинного обучения

44. Маркетинговый отдел имеет данные о 10 000 клиентов с 30 признаками (возраст, доход, частота покупок, категории товаров и т.д.). Необходимо разделить клиентов на однородные группы без использования размеченных данных для последующей персонализации предложений. Какой метод наиболее подходит для решения этой задачи?

45. Основное отличие модели множественной регрессии от модели парной регрессии заключается в том, что ...

46. Коэффициент корреляции Пирсона может принимать значения от ...

47. Установите соответствие между показателем и его эконометрической интерпретацией:

48. ... – это мера совместного изменения двух случайных величин, характеризующая направление их линейной связи

49. В модели множественной линейной регрессии $\hat{y} = 120 + 4x_1 - 2x_2$ (где $\hat{y}$ – прогнозное значение объема продаж (тыс. руб.); x_1 – расходы на рекламу (тыс. руб.); x_2 – цена товара (руб.)) коэффициент при переменной x_1 означает, что ...

50. Расположите в правильной последовательности этапы построения модели множественной регрессии:

51. Проблема, возникающая при сильной корреляции между объясняющими переменными, – это ...
52. ... – это метод оценки параметров линейной регрессионной модели, основанный на минимизации суммы квадратов отклонений фактических значений от расчетных
53. В числе последствий, которые может вызвать мультиколлинеарность в модели множественной регрессии, – ... (укажите 2 варианта ответа)
54. Для проверки общей статистической значимости модели множественной регрессии используется ...
55. При построении модели зависимости прибыли предприятия от затрат на рекламу, фонда оплаты труда и стоимости основных средств получены результаты: $R^2 = 0,89$. F-критерий статистически значим, однако два коэффициента при факторах оказались незначимыми по t-критерию, а между факторами наблюдается высокая корреляция. Какой вывод в этом случае является наиболее обоснованным?
56. Ключевыми признаками больших данных (Big Data) являются ...
57. К структурированным данным относятся ...
58. Установите соответствие между типом данных и его примером:
59. Хранилище данных Data Warehouse в отличие от озера данных Data Lake предназначено для ...
60. Классическая методика сбора данных, включающая извлечение данных из источников, их преобразование (очистку, нормализацию, обогащение) под нужную схему и последующую загрузку в хранилище данных, обозначается английской аббревиатурой ...
61. К основным источникам больших данных (Big Data) относят ... (укажите 4 варианта ответа)
62. Установите правильный порядок этапов обработки больших данных:
63. Основное отличие между языками запросов SQL и NoSQL заключается в том, что SQL ...
64. Технология, обеспечивающая распределенное хранение больших данных на кластерах серверов с использованием файловой системы HDFS, по-английски называется ...
65. Если в организации данные хранятся в сыром виде без предварительной очистки и схемы и могут быть использованы для любых типов анализа в будущем, то такая архитектура хранения называется ...
66. Расположите в правильной последовательности этапы ETL-процесса:
67. К методам обеспечения безопасности больших данных относятся такие, как ... (укажите 4 варианта ответа)

68. Программная среда R в первую очередь предназначена для ...
69. Одно из основных различий между языками R и Python заключается в том, что R ...
70. Установите соответствие между типом данных в R и его описанием:
71. Для импорта данных из CSV-файла в R используется функция ...
72. В языке R конструкция if-else используется для организации выполнения блоков кода в зависимости от выполнения логического ...
73. К базовым графическим возможностям языка R относятся такие функции, как ... (укажите 4 варианта ответа)
74. Установите правильный порядок этапов предварительной обработки данных:
75. Функция merge в R используется для ...
76. Цикл, который выполняет блок кода заранее известное количество раз, в R по-английски называется ...
77. Для расчета среднего арифметического значений числового вектора в R используется функция ...
78. Расположите в правильной последовательности этапы объединения двух таблиц данных с помощью оператора merge:
79. К функциям для расчета описательных статистик в R относятся такие, как ... (укажите 4 варианта ответа)
80. Алгоритм Apriori используется для ...
81. Минимальное значение поддержки (support) – это порог, задающий ...
82. Установите соответствие между понятием из области ассоциативных правил и его определением:
83. В методе ассоциативных правил правило вида $X \Rightarrow Y$ (где X и Y – это наборы элементов) интерпретируется таким образом, что ...
84. Анализ потребительской корзины является классическим примером применения метода ...
85. К сферам применения ассоциативных правил относятся такие, как ... (укажите 4 варианта ответа)
86. Установите правильный порядок работы алгоритма Apriori:
87. Правило будет считаться статистически значимым, если значение лифта lift ...

88. Метод анализа последовательностей знаков или событий, учитывающий порядок элементов, называется анализом последовательных ...
89. Информационная база для применения метода ассоциативных правил должна быть представлена в формате «...»
90. Расположите в правильной последовательности этапы применения ассоциативных правил в R:
91. К метрикам качества ассоциативных правил относятся такие, как ... (укажите 4 варианта ответа)
92. Метод k-средних относится к задачам ...
93. В методе опорных векторов (SVM) разделяющая гиперплоскость строится таким образом, чтобы ...
94. Установите соответствие между методом классификации и его характеристикой:
95. Функция `ram` из пакета `cluster` в R реализует такой алгоритм кластеризации, как ...
96. ... анализ – это совокупность методов многомерной классификации, позволяющих разбить множество объектов на однородные группы (кластеры) таким образом, чтобы объекты внутри кластера были более похожи друг на друга, чем на объекты из других кластеров
97. К итеративным методам кластерного анализа относятся такие, как ... (укажите 4 варианта ответа)
98. Установите правильный порядок применения метода k-средних в R:
99. Метод главных компонент PCA используется для ...
100. ... решений в задачах классификации представляет собой структуру, где каждый внутренний узел содержит проверку значения некоторого признака, ветвь – результат проверки, а лист – итоговый класс
101. Если при кластеризации алгоритм выбирает в качестве центров кластеров реально существующие объекты, а не средние значения, то такой метод называется ...
102. Расположите в правильной последовательности этапы построения случайного леса:
103. К методам снижения размерности относятся такие, как ... (укажите 4 варианта ответа)
104. Установите соответствие между значением коэффициента корреляции r и характеристикой корреляционной связи:
105. Графическое изображение зависимости между двумя количественными переменными называется полем ...
106. В ряду предпосылок классической линейной регрессионной модели – ... (укажите 3 варианта ответа)

- 107.** Определите правильную последовательность проведения парного корреляционно-регрессионного анализа:
- 108.** Если коэффициент корреляции r равен $-0,94$, то связь между переменными ...
- 109.** Если по данным выборки коэффициент корреляции r между расходами на рекламу и объемом продаж равен $0,8$, то коэффициент детерминации R^2 составит ... процента
- 110.** Если коэффициент детерминации равен $0,81$, то это означает, что ...
- 111.** Установите соответствие между показателем и его эконометрической интерпретацией:
- 112.** Обобщенная линейная модель множественной регрессии получается из классической модели при сохранении только ...
- 113.** Установите соответствие между нарушением предпосылок классической линейной модели множественной регрессии (КЛММР) и его следствием:
- 114.** Гетероскедастичность возникает, если ...
- 115.** Предпосылка классической линейной модели, требующая постоянства дисперсии случайной ошибки для всех наблюдений, – это ...
- 116.** К методам устранения гетероскедастичности относят ... (укажите 3 варианта ответа)
- 117.** ... регрессионная модель – это нелинейная модель по переменным, представляющая собой степенной ряд факторной переменной, остающаяся линейной по параметрам
- 118.** К линеаризуемым нелинейным моделям следует отнести ... (укажите 3 варианта ответа)
- 119.** Установите правильный порядок применения метода Бокса–Кокса:
- 120.** К функциям в R, являющимся инструментами для построения регрессионных моделей, относятся такие, как ... (укажите 4 варианта ответа)
- 121.** Компания внедряет систему сбора данных с тысяч датчиков интернета вещей (IoT-датчиков) на производственных линиях. Данные поступают непрерывно, имеют разный формат (температура, вибрация, давление, текстовые логи ошибок) и требуют мгновенной реакции при отклонениях. Руководство компании предлагает использовать традиционное реляционное хранилище (Data Warehouse) для накопления всей истории. Какая проблема возникнет в первую очередь при таком подходе?

- 122.** Компания внедряет систему сбора данных с датчиков на производственных линиях. Данные (температура, вибрация, текстовые логи ошибок) поступают непрерывно и требуют мгновенной реакции. Руководство компании решает использовать традиционную SQL-базу для хранения истории, но сталкивается с проблемами производительности. Какой из ключевых признаков больших данных был проигнорирован при выборе технологии?
- 123.** Аналитик получил задание проанализировать эффективность рекламной кампании. В его распоряжении оказались Excel-файл с расходами, JSON-логи с сайта, текстовые отзывы пользователей и видеозаписи с вебинаров. При попытке загрузить все в одну таблицу возникли ошибки, так как часть данных не имела четкой структуры и не помещалась в стандартные поля. Какие два типа данных оказались в этом наборе? В чем заключалась основная проблема при загрузке данных?
- 124.** Аналитик работает над проектом в программной среде R. Он написал код с использованием пакета dplyr. Коллега, запустив тот же код на своем компьютере с более новой версией dplyr, получил ошибку. Аналитик хочет, чтобы код воспроизводился на любом компьютере с одинаковыми версиями пакетов. Какой инструмент следует использовать для изоляции версий пакетов в проекте?
- 125.** Аналитик загрузил данные из CSV-файла командой `df <- read.csv("data.csv")` и получил предупреждение о том, что текстовые столбцы преобразованы в факторы. При построении модели это привело к неожиданному поведению модели. Какой параметр следовало указать в `read.csv()` при импорте, чтобы текстовые столбцы остались символьными?
- 126.** Аналитик написал несколько команд в консоли RStudio, построил график, создал переменные. Через час RStudio неожиданно закрылась, и весь код был потерян. Аналитику пришлось все восстанавливать. Какое правило организации работы в RStudio было нарушено?
- 127.** В маркетинговом отделе розничной сети провели анализ покупательских корзин и получили ассоциативное правило: {хлеб, масло} ⇒ {молоко}. Известно, что поддержка (support) этого правила составляет 0,3, а достоверность (confidence) – 0,8. Как интерпретируется достоверность данного правила?
- 128.** При поиске ассоциативных правил в данных о покупках был найден набор правил. Для одного из правил метрика Lift (лифт) оказалась равна 0,5. О чем свидетельствует это значение?
- 129.** Для поиска ассоциативных правил в больших базах транзакций часто используют алгоритм Apriori. В основе его работы лежит свойство, которое позволяет значительно сократить количество проверяемых наборов товаров. В чем заключается ключевое свойство, используемое алгоритмом Apriori?
- 130.** Аналитик работает с многомерным набором данных, содержащим 50 коррелирующих признаков. Он хочет уменьшить размерность данных, сохранив при этом максимум информации, и планирует использовать метод главных компонент (PCA). Какое утверждение о методе PCA является верным?

- 131.** Компания провела кластеризацию клиентской базы методом k-средних и получила 5 сегментов. Для оценки качества разбиения клиентской базы на кластеры аналитик рассчитал коэффициент силуэта для каждого значения k и получил следующие результаты: $k = 3 \rightarrow 0.65$, $k = 4 \rightarrow 0.72$, $k = 5 \rightarrow 0.58$, $k = 6 \rightarrow 0.51$. Какое количество кластеров следует выбрать исходя из результатов анализа силуэта?
- 132.** Специалист по машинному обучению обучает модель классификации для выявления мошеннических транзакций. На обучающей выборке модель демонстрирует точность 99 %, однако на тестовой выборке точность падает до 72 %. При этом модель содержит несколько сотен признаков и использует полиномы высокой степени. Какое явление наиболее вероятно привело к данной ситуации и какой метод из темы «Снижение размерности» поможет ее решить?
- 133.** Аналитик построил модель линейной регрессии для прогноза продаж и получил коэффициент детерминации $R^2 = 0,95$ на обучающей выборке. Однако при проверке модели на тестовой выборке R^2 упал до 0,32. Модель демонстрирует высокую точность на данных, на которых обучалась, и низкую – на новых данных. Какой вывод в этом случае наиболее обоснован?
- 134.** При прогнозировании спроса на товар построена регрессионная модель: $\hat{y} = 10 + 0,5x_1 - 0,2x_2$, где x_1 – цена конкурента, x_2 – собственная цена. Коэффициент регрессии при x_2 отрицательный, что соответствует экономической теории. Однако на практике при снижении собственной цены спрос не увеличивается. Анализ показал, что x_1 и x_2 изменяются синхронно (конкуренты следуют за ценовыми изменениями). Какое явление наиболее вероятно исказило оценку влияния собственной цены на спрос?
- 135.** При построении дерева решений для прогнозирования стоимости недвижимости алгоритм на каждом шаге выбирает признак и порог, которые обеспечивают максимальное снижение среднеквадратичной ошибки (MSE) в дочерних узлах. После остановки роста дерева была зафиксирована высокая ошибка на тестовой выборке, при том что на обучающей выборке ошибка была близка к нулю. Какая стратегия в первую очередь позволит повысить обобщающую способность модели?