

1. Английская аббревиатура ... обозначает диаграмму, отображающую причинно-следственные связи между различными факторами, действиями и результатами, которая используется для проектирования, анализа и оптимизации сложных решений в бизнесе и других областях
2. ... представляют собой одиночные текстовые данные (например, «А», «b», «1»)
3. Установите соответствие универсальных категорий типы данных и их характеристик:
4. Расположите характеристики неструктурированных, полуструктурированных и структурированных данных в порядке перечисления этих типов данных в задании:
5. ... в программировании – это структура данных, которая представляет собой фиксированный по размеру упорядоченный набор элементов одного типа с доступом по числовому индексу
6. ... тип данных в контексте теории баз данных – это атрибут (или набор атрибутов), значения которого однозначно определяют каждый экземпляр сущности (каждую строку в таблице)
7. ... – это технология непрерывной обработки данных в режиме реального времени по мере их поступления от источников (потоков данных), без предварительного сохранения на диск
8. ... данных – это последовательность этапов, через которые проходят данные от момента их создания или получения до окончательного архивирования или удаления
9. ... – это криптографический процесс преобразования исходных данных (открытого текста) в зашифрованную форму (шифртекст) с использованием специального алгоритма и ключа
10. ... ключ – это искусственный уникальный идентификатор, который добавляется в таблицу базы данных исключительно с целью идентификации записей и не имеет никакого смыслового (бизнес) значения
11. Хранилища данных могут быть ..., если они включают в себя такие системы, как NoSQL, например, колоночные или графовые базы данных
12. ... – это программная система для централизованного хранения и управления большим объемом структурированных данных, собранных из разных источников, чтобы использовать их в аналитике и отчетности, прогнозировании и принятии бизнес-решений
13. Data ... – это централизованное хранилище всевозможных типов данных, в котором можно хранить большие объемы данных в их исходном формате, без преобразований

14. Установите соответствие систем управления базами данных (СУБД), классифицированных по модели представления данных, и их примеров:
15. Системы управления базами данных (СУДБ) класса NoSQL, которые служат для представление данных в виде узлов и связей – это ... NoSQL-СУДБ
16. Английская аббревиатура ... обозначает операционные системы управления базами данных
17. ... – это метод горизонтального масштабирования базы данных, при котором данные разделяются на независимые части и хранятся на разных серверах; увеличивает производительность и емкость системы
18. ... – это логическая единица работы с базой данных, состоящая из набора операций, которые должны быть выполнены как единое целое; обеспечивает целостность данных
19. ... файловая система – это файловая система, позволяющая получить доступ к файлам с множества хостов через компьютерную сеть; обеспечивает совместный доступ к файлам и прозрачную работу с ними
20. Расположите характеристики фреймворков Apache Atlas, Apache Flink, Apache Spark в порядке их перечисления в задании:
21. Установите соответствие языков программирования и их библиотек:
22. ... – это разработанная компанией JetBrains популярная интегрированная среда разработки, которая специализируется на языке Python; предлагает много функций: интеллектуальное автодополнение кода; отладчик; поддержка виртуальных окружений; инструменты для тестирования; интеграция с системами управления версиями
23. Установите соответствие ключевых библиотек Python и их характеристик:
24. Корреляционная ... – это квадратная таблица, в которой указаны коэффициенты корреляции между всеми парами числовых переменных; используется для анализа мультиколлинеарности и отбора признаков
25. ... вариации – это простая мера вариации, вычисляемая как разница между максимальным и минимальным значениями
26. ... – это визуализация числовых значений в виде цветовой шкалы; применяется для анализа корреляций, распределений или матриц ошибок, где интенсивность цвета отражает величину показателя
27. Расположите описания критериев Андерсона–Дарлинга, Колмогорова–Смирнова и Шапиро–Уилка в порядке их перечисления в задании:

28. ... – это визуализация долей категориальных переменных в виде секторов круга, где каждый сектор отображает относительную долю категории в процентах
29. Установите соответствие разных типов графиков и их назначения:
30. ... данных – это масштабирование числовых переменных в заданный диапазон, обычно $[0, 1]$; используется для ускорения сходимости моделей и предотвращения доминирования признаков с большими значениями
31. Термин «... размерности» ввел в 1961 г. Ричард Беллман, который описывает совокупность явлений, возникающих при анализе данных в пространствах с высокой размерностью (неблагоприятные последствия, возникающие при работе с многомерными данными)
32. Установите соответствие понятий и их определений:
33. Установите соответствие основных целей уменьшения размерности и их описаний:
34. Установите правильный порядок шагов алгоритма реализации метода главных компонент (PCA) в Python:
35. ... нагрузки – это коэффициенты, показывающие, насколько сильно каждая переменная связана с определенным фактором
36. ... – это ситуация в машинном обучении, когда модель слишком точно подстраивается под обучающие данные и не просто учит закономерности, а запоминает каждый пример, включая ошибки, шум и случайные отклонения
37. ... – это линейный метод построения новых признаков, который находит новые оси (компоненты), вдоль которых дисперсия данных максимальна; основная цель этого метода – уменьшить размерность при минимальной потере информации
38. Перед применением метода главных компонент (PCA) необходима ... данных – приведение каждого признака к общему масштабу с нулевым средним и единичным стандартным отклонением
39. Одним из подходов к снижению размерности является построение новых признаков (feature extraction) – к методам этого подхода относятся ... (укажите 2 варианта ответа)
40. ... – это мера разброса данных, представляющая собой средний квадрат отклонений наблюдений от среднего значения
41. ... анализ – это статистический метод многомерной обработки данных, задачей которого является группировка совокупности объектов в однородные группы на основе их схожести

42. ... – это состояние, при котором модель слишком проста и не способна уловить закономерности в данных, что ведет к низкой точности
43. Расположите описания таких метрик расстояния между объектами в пространстве признаков, как евклидово расстояние, манхэттенское расстояние и косинусное расстояние, в порядке их перечисления в задании:
44. ... – это один из самых распространенных алгоритмов кластеризации, который обнаруживает кластеры произвольной формы и помечает выбросы, но чувствителен к параметрам
45. ... – это мера неопределенности выборки, применяемая в деревьях решений; чем меньше ее значение, тем более однороден узел, а если все объекты принадлежат одному классу, ее значение равно нулю
46. Установите соответствие этапов кластерного анализа и их содержания:
47. Метод ... – это метод определения числа кластеров (для K-means) в алгоритмах кластеризации, который анализирует, как уменьшается внутрикластерная дисперсия с ростом числа k ; искомым k является точка на графике, после которой дисперсия снижает темп своего падения
48. Стандартизация (Standardization) – это ...
49. ... – это целевая переменная в машинном обучении, значение которой модель должна предсказать или объяснить на основе других данных
50. ... в машинном обучении – это добавление штрафа за сложность модели к функции потерь
51. ... множество – это набор элементов, который встречается в транзакционной базе данных чаще или равно установленному порогу поддержки; такие множества лежат в основе генерации ассоциативных правил
52. Набор элементов, которые приобретены или встречаются вместе в одной единице наблюдения, – это ...
53. ... – это совокупность элементов, встречающихся совместно в одной транзакции
54. ... времени – это заданный временной диапазон, в пределах которого рассматриваются последовательности событий при анализе шаблонов; его ограничение помогает исключать нерелевантные связи между событиями, происходящими слишком далеко друг от друга
55. ... – это операция, которая позволяет найти элементы, которые присутствуют одновременно в нескольких транзакциях, помеченных идентификаторами

56. ... представление данных – это способ хранения транзакционных данных, при котором каждому элементу сопоставляется список идентификаторов транзакций (TID-list), в которых он встречается; такой формат ускоряет вычисления поддержки за счет операций пересечения списков
57. Установите соответствие названий и описаний алгоритмов поиска ассоциативных правил и последовательных шаблонов, решающих задачу поиска часто встречающихся наборов элементов и генерации правил на их основе:
58. Установите правильный порядок протекания процесса генерации обобщенных правил:
59. ... – это минимальная доля транзакций, в которых встречается определенный набор элементов; используется для фильтрации нечастых комбинаций и снижения количества кандидатов
60. Алгоритм ... не генерирует все шаблоны, а рекурсивно строит их, создает из часто встречающихся префиксов (аналог FP-Growth, но для последовательностей)
61. ... – это внутреннее кодирование решения в генетических алгоритмах (например, список параметров), с которым работают операторы мутации и скрещивания
62. ... в генетических алгоритмах – это реальное выражение генотипа, интерпретируемое как готовое решение задачи (например, структура модели или набор признаков)
63. ... в контексте генетических алгоритмов – это полный цикл эволюции популяции, включающий отбор, скрещивание, мутацию и формирование нового набора решений
64. Установите соответствие понятий и их определений:
65. Расположите (в порядке перечисления в задании) характеристики таких категорий методов выбора оптимальных предикторов при построении моделей машинного обучения, как методы-обертки, встроенные методы и методы-фильтры:
66. ... операторы – это набор процедур в генетических алгоритмах (мутация, скрещивание, селекция), которые управляют поиском оптимальных решений
67. ... интервалы – это диапазоны значений оцененных параметров, которые с заданной вероятностью (обычно 95 %) содержат истинное значение параметра, отражая неопределенность оценивания
68. ... – это метод-обертка в машинном обучении, цель которого состоит в том, чтобы уменьшить размерность входного пространства и при этом сохранить максимально возможную предсказательную способность модели
69. Английская аббревиатура ... обозначает метод интерпретации моделей, основанный на теории кооперативных игр Шепли, который распределяет вклад каждого признака в итоговое предсказание модели

- 70.** ... – это нарушение предположения о постоянстве дисперсии ошибки регрессионной модели, при котором разброс остатков зависит от значений объясняющих переменных; приводит к некорректным стандартным ошибкам
- 71.** Индекс ... – это показатель степени логичности оценок в матрице парных сравнений; показывает, насколько заданные экспертные суждения близки к внутренне непротиворечивой системе предпочтений
- 72.** Установите соответствие понятий и их значений:
- 73.** Установите соответствие признаков классификации и соответствующих классов задач принятия решений:
- 74.** Иерархическая ... – это естественный способ упростить сложные задачи, который предлагает метод анализа иерархий (МАИ); главная идея здесь заключается в том, что любую сложную задачу можно разложить на более простые составляющие, а затем управлять ими последовательно, уровень за уровнем
- 75.** ... – это количественная мера субъективной ценности результата для лица, принимающего решение; позволяет сравнивать альтернативы в условиях риска и неопределенности
- 76.** Класс задач принятия решений, характерных для финансовых решений (инвестиции, портфельные стратегии), страхования, технического контроля и маркетинговых прогнозов, – это задачи принятия решений в условиях ...
- 77.** ... неопределенность – это форма описания неопределенности, при которой параметры задаются не точными значениями, а диапазонами; требует специальных методов анализа, учитывающих возможные вариации внутри интервала
- 78.** Расположите описания главных свойств иерархии в методе анализа иерархий (МАИ) – таких как ступенчатость, декомпозируемость, однозначность состава уровней, завершенность, независимость уровней – в порядке их перечисления в задании:
- 79.** ... решение по Парето – это альтернатива, которую нельзя улучшить по одному критерию без ухудшения по другому; формирует основу выбора в многокритериальной оптимизации
- 80.** ... оценки – это суждения специалистов, используемые при недостатке данных; применяются для задания критериев, сравнений и параметров моделей, требуют проверки согласованности и обоснованности
- 81.** Английская аббревиатура ... обозначает набор протоколов, инструментов и определений для взаимодействия между различными программными приложениями; позволяет различным системам обмениваться данными и функциональностью стандартизированным способом
- 82.** ... – это наборы данных, которые настолько велики или сложны, что традиционные методы обработки данных не справляются с ними

83. Установите соответствие универсальных категорий типы данных и их характеристик:

84. Установите соответствие типов данных и их характеристик:

85. ... хранятся в реляционных базах данных (SQL) и управляются с помощью языка запросов SQL

86. ... – это структура данных, которая представляет собой фиксированный по размеру упорядоченный набор элементов одного типа с доступом по числовому индексу

87. Установите соответствие понятий и их содержания:

88. Расположите расширения, используемые для сохранения таких неструктурированных данных, как текстовые документы, изображения, аудиофайлы и видеофайлы, в порядке их перечисления в задании:

89. ... – это процесс автоматизированного сбора, анализа и преобразования данных, как правило, из неструктурированных или полуструктурированных источников в структурированный формат, пригодный для дальнейшего анализа и хранения

90. Первичный ... – это уникальный идентификатор каждой записи в таблице реляционной базы данных

91. Установите правильный порядок работы с большими данными и принятия решений на их основе:

92. Хранилище данных, где данные организованы в таблицы, состоящие из строк и колонок, принято называть ...; такая модель хорошо подходит для структурированных данных и использует широко распространенный стандартный язык SQL

93. ... данных (Data Warehouse, DWH) – это программная система для централизованного хранения и управления большим объемом структурированных данных, собранных из разных источников, чтобы использовать их в аналитике и отчетности, прогнозировании и принятии бизнес-решений

94. ... – это централизованное хранилище всевозможных типов данных, в котором можно хранить большие объемы данных в их исходном формате, без преобразований

95. Установите соответствие систем управления базами данных (СУБД), классифицированных по модели представления данных, и их описаний:

96. ... – это процесс копирования и поддержания согласованности данных между несколькими экземплярами базы данных; обеспечивает отказоустойчивость, доступность и распределение нагрузки

97. ... система управления базами данных (СУБД) – это СУДБ, компоненты которой расположены на различных узлах компьютерной сети; обеспечивает прозрачный доступ к данным как к единому целому

98. Установите соответствие систем управления базами данных СУДБ класса NoSQL и их примеров:

99. Английская аббревиатура OLTP обозначает ... системы управления базами данных

100. Установите соответствие фреймворка и его характеристики:

101. Установите соответствие языков программирования и их характеристики:

102. ... в статистике – это середина упорядоченной выборки данных, значение, делящее упорядоченные данные пополам

103. Расположите описания ключевых библиотек Python, таких как NumPy, Pandas и Matplotlib, в порядке их перечисления в задании:

104. ... – это мера линейной зависимости между двумя переменными

105. ... – это визуализация числовых значений в виде цветовой шкалы; применяется для анализа корреляций, распределений или матриц ошибок, где интенсивность цвета отражает величину показателя

106. Установите соответствие тестов и их описаний:

107. Критерий ... – это непараметрический аналог и альтернатива для дисперсионного анализа (ANOVA); применяется для сравнения трех и более групп при отсутствии нормального распределения; основан на анализе ранговых сумм

108. ... – это графическое представление частотного распределения числовых данных, где ось X отображает интервалы значений, а ось Y – количество наблюдений в каждом интервале

109. Расположите задачи, решаемые с помощью гистограммы, диаграммы рассеяния, линейного графика и boxplot (ящика с усами), в порядке перечисления этих видов графика в задании:

110. Термин «проклятие ...» ввел в 1961 г. Ричард Беллман, который описывает совокупность явлений, возникающих при анализе данных в пространствах с высокой размерностью (неблагоприятные последствия, возникающие при работе с многомерными данными)

111. ... – это метод визуализации, показывающий одновременно наблюдения и признаки в пространстве главных компонент; точки обозначают объекты, а векторы – признаки, что помогает анализировать взаимосвязи между ними и выявлять направления наибольшего вклада переменных

112. ... – это тип нейронной сети, обучаемой восстанавливать исходные данные из их сжатого представления; архитектура включает кодировщик (encoder), преобразующий данные во внутреннее пространство меньшей размерности, и декодировщик (decoder), восстанавливающий данные обратно

113. Установите соответствие линейных методов построения новых признаков и описаний этих методов:

114. Алгоритм машинного обучения для визуализации ... относится к нелинейным методам построения новых признаков, это визуализационный метод, который сохраняет локальные расстояния между точками, хорошо показывает кластеры, но плохо работает на больших выборках

115. Одним из подходов к снижению размерности является отбор признаков (feature selection) – к методам этого подхода относятся ... (укажите 2 варианта ответа)

116. ... вращение – это метод вращения факторов, допускающий коррелированные факторы, что повышает реалистичность интерпретации данных по сравнению с ортогональным вращением

117. ... – это метод оценки устойчивости и обобщающей способности модели, при котором данные многократно разбиваются на обучающие и тестовые подмножества

118. Установите правильную последовательность этапов кластерного анализа:

119. Установите соответствие метрик расстояния между объектами в пространстве признаков и описаний этих метрик:

120. ... – это один из самых распространенных алгоритмов кластеризации, который является быстрым и простым, но требует заранее задать число кластеров и находит только сферические группы

121. ... – это метод определения числа кластеров (для K-means) в алгоритмах кластеризации, который анализирует, как уменьшается внутрикластерная дисперсия с ростом числа k ; искомым k является точка на графике, после которой дисперсия снижает темп своего падения

122. Английская аббревиатура ... обозначает метрику, измеряющую площадь под ROC-кривой и показывающую способность модели различать положительные и отрицательные классы; чем ближе ее значение к 1, тем лучше классификатор отделяет классы

123. Установите соответствие понятий и их значений:

124. ... – это параметр, задаваемый до обучения модели (например, количество деревьев, скорость обучения); контролирует процесс обучения, но не подбирается из данных напрямую

125. Базовая единица анализа в транзакционной базе, например, отдельный товар, действие пользователя или событие, – это ...

126. ... – это набор элементов, которые приобретены или встречаются вместе в одной единице наблюдения

- 127.** ... элементов – это совокупность элементов, встречающихся совместно в одной транзакции (например, его образуют элементы {молоко, хлеб, масло}, если покупатель приобрел их все вместе)
- 128.** ... – это заданный временной диапазон, в пределах которого рассматриваются последовательности событий при анализе шаблонов; его ограничение помогает исключать нерелевантные связи между событиями, происходящими слишком далеко друг от друга
- 129.** ... – это визуальная схема или план расположения товаров в торговом пространстве, основанная на анализе ассоциативных правил; используется для оптимизации выкладки, чтобы повысить кросс-продажи
- 130.** ... представление данных – это способ хранения транзакций, при котором каждая запись представляет одну транзакцию с перечнем всех входящих в нее элементов; такой формат удобен для первоначального сбора данных, но менее эффективен при вычислениях частоты
- 131.** Установите правильный порядок основных шагов алгоритма Eclat:
- 132.** Установите соответствие основных алгоритмов поиска последовательных шаблонов и характеристик данных алгоритмов:
- 133.** Расположите определения понятий «статистическая значимость», «функция приспособленности» и «невидимая гетерогенность» в порядке их перечисления в задании:
- 134.** Установите соответствие категорий методов выбора оптимальных предикторов при построении моделей машинного обучения и описаний данных методов:
- 135.** ... алгоритмы – это класс эволюционных методов оптимизации и поиска, вдохновленных процессами естественного отбора и биологической эволюции
- 136.** Суть метода-фильтра ... для отбора признаков в машинном обучении с использованием – в том, что признаки, которые не прошли порог значимости, удаляют
- 137.** Установите соответствие методов-оберток в машинном обучении и их характеристик:
- 138.** Установите правильный порядок шагов алгоритма рекурсивного исключения признаков (RFE)
- 139.** Английская аббревиатура ... обозначает метод локальной интерпретации сложных моделей, который аппроксимирует поведение модели простой интерпретируемой моделью в окрестности конкретного наблюдения
- 140.** ... предпочтений – это процедура объединения индивидуальных предпочтений нескольких экспертов в одно согласованное коллективное мнение; используется в групповых решениях, требует учета согласованности, влияния отдельных участников и устойчивости к манипуляциям

- 141.** Расположите характеристики понятий «аддитивная функция полезности», «лексикографический порядок» и «Парето-доминирование» в порядке их перечисления в задании:
- 142.** ... – это количественная мера субъективной ценности результата для лица, принимающего решение; позволяет сравнивать альтернативы в условиях риска и неопределенности
- 143.** Установите соответствие классов задач принятия решений и их характеристик:
- 144.** Класс задач принятия решений, характерных для сфер, где поведение системы поддается строгому математическому описанию (например, оптимизация ресурсов, линейное программирование, распределение производственных мощностей), – это задачи принятия решений в условиях ...
- 145.** ... парных сравнений – это таблица, в которой элементы сравниваются по два, отражая относительную важность одного элемента над другим; используется в методе анализа иерархий (АИИ) для вычисления весов критериев и оценки согласованности суждений
- 146.** Установите соответствие свойств иерархии в методе анализа иерархий (МАИ) и их описаний:
- 147.** Томас Саати предложил для метода анализа иерархий (МАИ) целочисленную ... шкалу, которая стала стандартом в методах принятия решений
- 148.** ... алгоритм многокритериальной оптимизации (NSGA-II) – это метод, имитирующий эволюцию популяции решений для поиска множества недоминантных альтернатив; эффективен в сложных нелинейных задачах
- 149.** Построение системы принятия решений на основе данных невозможно без создания соответствующих организационных условий. Здесь выделяют три ключевых компонента. Говоря об одном из этих компонентов, можно утверждать, что речь здесь о создании архитектуры, которая одновременно отвечает потребностям бизнеса и требованиям регулирования. Назовите этот компонент
- 150.** Можно выделить такие фундаментальные этапы аналитического процесса, как: сбор и интеграция данных, очистка и подготовка данных; анализ и моделирование; визуализация и интерпретация; принятие решений; оптимизация и выбор подхода; исполнение и ретроспектива. Разберем подробнее один из этапов. Сырые данные часто содержат ошибки, дубли и пробелы. Данный этап включает: удаление дублей; нормализацию форматов (например, единые валюты и часовые пояса); работу с пропущенными значениями; трансформацию к унифицированным схемам. На этом этапе важно учитывать и метаданные – описание происхождения и контекста данных, без которых последующий анализ может исказить результаты. Назовите описанный этап.
- 151.** Можно выделить такие фундаментальные этапы аналитического процесса, как: сбор и интеграция данных, очистка и подготовка данных; анализ и моделирование; визуализация и интерпретация; принятие решений;

оптимизация и выбор подхода; исполнение и ретроспектива. Разберем подробнее один из этапов. На этом этапе решение запускают в работу, но процесс на этом не заканчивается. Важно организовать: мониторинг ключевых показателей и достижения целей; анализ качества самого процесса принятия решений, а не только результатов; хранение и документирование решений для повторного использования. Назовите описанный этап.

152. Несколько преимуществ, которые бизнес может получить, если он имеет грамотно спроектированную систему DWH (хранилище данных). Например, DWH в зависимости от изначальных требований может содержать всю историю данных. При этом аналитику, который хочет получить отчет с метриками за прошлые годы, не придется беспокоиться о том, что при запуске отчета не найдется данных. О каком преимуществе для компаний говорится в данном примере?

153. Системы управления базами данных (СУБД) могут быть классифицированы по архитектуре. Так, есть СУБД, где база данных хранится на множестве серверов (узлов), объединенных в кластер. Это позволяет обрабатывать огромные объемы информации. О каких СУБД идет речь?

154. Благодаря внедрению YARN в 2013 г. фреймворк Hadoop превратился из решения по пакетной обработке данных в опорную платформу для нескольких других систем программирования. Примером может служить инструмент для выполнения запросов к крупным наборам данных, хранящимся в системе Hadoop. О каком инструменте идет речь?

155. Данный язык программирования является относительно новым. Он становится все популярнее среди специалистов по анализу данных. Одна из ключевых характеристик этого языка – его высокая производительность. Он разработан так, чтобы позволить эффективно выполнять вычисления и анализировать большие объемы данных. Этот язык использует современные техники компиляции и оптимизации, что делает его почти таким же быстрым, как низкоуровневые компилируемые языки, такие как C и Fortran. Данный язык идеально подходит для анализа данных и научных расчетов. С его помощью можно легко загрузить, обработать и проанализировать большие наборы данных. Этот язык уже имеет огромное количество доступных пакетов, что позволяет аналитикам и исследователям легко выполнять разные задачи, от статистического анализа и машинного обучения до численных расчетов и визуализации данных. Назовите описанный язык программирования.

156. Данный язык программирования является мощным инструментом для анализа данных и статистики. Его часто используют в корпоративной среде для принятия стратегических решений. Это язык предоставляет обширный функционал для выполнения разных задач, от базовой статистики до сложных бизнес-аналитических задач. Он предоставляет богатый набор инструментов для обработки данных, проведения статистических анализов, создания прогнозов, моделей и визуализации результатов. Назовите описанный язык программирования.

157. Данный язык программирования предоставляет мощные инструменты для численных расчетов и анализа данных. Он популярен в академической и инженерной среде. Одна из ключевых характеристик этого языка – его поддержка матричных вычислений. Это делает его особенно ценным для инженерных и научных расчетов, где

матрицы часто используют для представления данных и моделей. Данный язык используют для обучения студентов и проведения исследований в разных областях, в том числе в сфере физики, инженерии, экономики и биологии. Назовите описанный язык программирования.

158. После снижения размерности важно оценить, насколько новая система признаков сохранила информацию исходных данных. Цель – найти проекцию, при которой потери информации минимальны. Для каких методов информативность выражается через сохранение расстояний или плотностей между объектами в новом пространстве?

159. Существует нелинейный метод снижения размерности, сохраняющий локальную структуру данных за счет аппроксимации каждой точки линейной комбинацией ее ближайших соседей. Назовите этот метод.

160. После снижения размерности важно оценить, насколько новая система признаков сохранила информацию исходных данных. Нужно найти проекцию, при которой потери информации минимальны. Выделяют формальные меры информативности. К ним относится метод, который измеряет степень разделимости классов в новом пространстве признаков. Его цель – найти линейную комбинацию признаков, которая лучше всего разделяет два или более классов данных. Назовите описанный метод.

161. Существует параметрический алгоритм машинного обучения, который используют для решения задач классификации и регрессии. Он формирует модель в виде иерархической структуры, где каждый внутренний узел соответствует проверке условия на одном из признаков, каждая ветвь – это результат проверки условия, а лист – метка класса (в классификации) или числовое значение (в регрессии). Задача алгоритма заключается в том, чтобы найти последовательность логических правил, которая делит пространство признаков на подмножества, в каждом из которых целевая переменная максимально однородна. Назовите описанный алгоритм.

162. Случайный лес (Random Forest Classifier) – один из самых популярных и мощных алгоритмов машинного обучения для классификации и регрессии. Вместо того чтобы строить одно большое дерево, случайный лес строит множество небольших деревьев, каждое из которых принимает собственное решение. Затем все деревья объединяют, и итоговое решение определяют большинством голосов. Такой ансамбль работает по принципу коллективного интеллекта: если одно дерево ошибается, другие деревья могут компенсировать эту ошибку. На одном из этапов работы алгоритм при построении дерева в каждом узле выбирает случайное подмножество признаков. Это заставляет деревья «думать по-разному» и снижает корреляцию между ними. Назовите описанный шаг работы алгоритма Случайный лес.

163. Существует показатель неоднородности выборки в узле дерева решений. Чем меньше его значение, тем «чище» узел. Данный критерий используется при построении деревьев решений для выбора оптимального разделения данных. Назовите описанный критерий.

- 164.** Эти правила являются инструментом в области анализа данных и интеллектуального анализа больших массивов информации. Они позволяют выявлять закономерности и взаимосвязи между объектами в наборах данных, которые могут быть неочевидными при поверхностном рассмотрении. О каких правилах говорится в описании?
- 165.** Чтобы оценивать полезность и значимость ассоциативных правил, применяют три ключевые метрики. Одна из этих метрик – это вероятность того, что В встречается в транзакции, если в ней уже присутствует А (где А и В – множества элементов или событий, которые встречаются в транзакциях). Назовите данную метрику.
- 166.** Чтобы оценивать полезность и значимость ассоциативных правил, применяют три ключевые метрики. Одна из этих метрик показывает, насколько вероятность появления В увеличивается при наличии А по сравнению с вероятностью В в целом. То есть метрика показывает, насколько сильнее события А и В встречаются вместе, чем можно было бы ожидать, если бы они были независимы. Назовите данную метрику.
- 167.** Основная идея данного метода заключается в том, что модель обучается на всех доступных признаках, после чего на основе важности признаков наиболее слабый (наименее информативный) удаляется. Это продолжается до тех пор, пока не останется заданное количество признаков. Назовите описанный метод.
- 168.** Существует AutoML-библиотека в Python, которая использует генетические алгоритмы для построения оптимального конвейера машинного обучения (ML-пайплайна). Эта библиотека может: выбирать и комбинировать алгоритмы машинного обучения; автоматически подбирать гиперпараметры; выполнять отбор признаков и др. Назовите описанную библиотеку.
- 169.** Существует высокоуровневый фреймворк в Python, предназначенный для проектирования, настройки и выполнения эволюционных вычислений. Он создан для исследователей, специалистов по оптимизации и разработчиков, которым нужно быстро прототипировать генетические и эволюционные методы. Данный фреймворк сочетает гибкость, модульность и богатую библиотеку операторов, позволяет строить как классические генетические алгоритмы, так и сложные эволюционные стратегии, генетическое программирование и гибридные схемы. Назовите описанный фреймворк.
- 170.** Задачи принятия решений по степени определенности информации о среде и последствиях выбора классифицируют на задачи принятия решений в условиях риска, в условиях неопределенности и в условиях определенности. Задачи одного из этих классов называют задачами с неполной или неформализуемой информацией. В рамках неопределенности применяют не вероятностные, а критериальные правила (например, используют критерий Вальда, критерий Сэвиджа, критерий Гурвица). Этот раздел классификации наиболее тесно связан с реальными управленческими, политическими и стратегическими решениями. Математический аппарат здесь не может заменить весь процесс рассуждений и служит лишь инструментом поддержки. Назовите описанный класс классификации задач принятия решений.

171. В рамках неопределенности применяют не вероятностные, а критериальные правила. Один из критериев, используемых в принятии решений в условиях неопределенности, дает сбалансированный подход между крайним оптимизмом и крайним пессимизмом, предлагая учитывать одновременно наихудший возможный результат (пессимистический подход) и наилучший возможный результат (оптимистический подход). Назовите этот критерий.

172. Иерархия в методе анализа иерархий (МАИ) – это формализованная модель задачи принятия решения, отражающая связь общей цели с факторами, которые влияют на ее достижение, структуру критериев и их подчиненность, набор альтернатив, из которых выбирают решение. Чтобы иерархия работала корректно и отражала реальную логику лица, принимающего решение, важно соблюдать несколько методологических правил. Согласно одному из этих правил, нельзя использовать дублирующие критерии. Например, нельзя одновременно использовать критерии «стоимость владения» и «стоимость обслуживания», если первое уже включает второе. Назовите описанное методологическое правило.