МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»
Обнинский институт атомной энергетики –
филиал федерального государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский ядерный университет «МИФИ»
(ИАТЭ НИЯУ МИФИ)

ОТДЕЛЕНИЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ НАУК (О)

Одобрено на заседании
Ученого совета
ИАТЭ НИЯУ МИФИ
Протокол №25.8 от 25.08.2025

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ
для студентов.
Терминологический словарь по дисциплине

АНАЛИЗ БОЛЬШИХ ДАННЫХ

название дисциплины

для студентов направления подготовки

38.03.05 Бизнес-информатика

код и название направления подготовки

образовательная программа

Цифровая трансформация компании

Форма обучения: очная

г. Обнинск 2025 г.

ВВЕДЕНИЕ

Терминологический словарь по дисциплине «Анализ больших данных» способствует систематизации знаний студентов ввиду активизации их самостоятельной работы с базой источников, а именно, с нормативно-правовыми актами, специальной литературой, электронными ресурсами.

В состав словаря включены специальные слова и значения, которые являются узкопрофессиональными терминами по данной дисциплине.

Значение термина раскрывается в кратком определении, достаточном для понимания самого слова и его употребления.

Терминологический словарь не содержит сведения для всестороннего знакомства с самим называемым определением.

Краткий терминологический словарь

1. **Авторизация** (Authorization) - разрешение на доступ к ресурсам или службам.
2. **Алгоритм градиентного бустинга** (boosting — улучшение, xgboost) — процедура последовательного построения композиции алгоритмов машинного обучения с целью улучшения качества классификации или предсказательной силы модели.
3. **Асимметричное распределение** — асимметричное распределение данных имеет длинный хвост справа с несколькими высокими значениями (положительно скошенное) или длинный хвост слева с несколькими низкими значениями (отрицательно скошенное).
4. **Аутентификация** (Authentication) - установление подлинности. Процесс идентификации участвующей в соединении стороны.
5. **База данных** (Database) - организованный массив структурированных данных.
6. **Байесовский метод вывода** — вывод на основе теоремы Байеса, использует не только текущую информацию, но и прежнее суждение о гипотезе для оценки апостериорной вероятности, оценивающей уровень доверия к гипотезе после наблюдаемых событий.
7. **Большие данные** (Big Data) — включает в себя стратегии, технологии и информационные системы, направленные на получение, обработку, хранение, анализ и визуализацию сложных структурированных и неструктурированных наборов данных с помощью пакетной обработки, потоковой обработки, NoSQL, HPC, MPP, In-Memory и других.
8. **Вариация остатков** — вариация переменной, которая остаётся после того, как удалена вариабельность, относящаяся к интересующим нас факторам. Это вариация, не объяснимая моделью, также называется «ошибочная, или необъяснённая, вариация».
9. **Вероятностная модель** — математическое представление случайного события. Определяется пространством элементарных событий и вероятностью событий.
10. **Выборка обучающая (training sample)** — выборка, на которой производится обучение алгоритма, в частности, нейронной сети с целью минимизации заданной функции потерь.
11. **Выборка проверочная (validation sample)** — выборка, на которой осуществляется проверка модели из множества моделей, построенных по обучающей выборке и выбирается лучшая модель.
12. **Диаграмма «стебель-листья»** — полуграфический метод, используемый для представления числовых данных, в котором первая (крайняя слева) цифра каждого значения данных является стеблем, а остальные цифры числа — это листья.
13. **Интерквартильный размах** — разница между первым и третьим квартилем (IQR).
14. **Клетка таблицы сопряжённости** — пересечение отдельной строки и отдельного столбца таблицы сопряженности. Матрица ошибочной класси-

фикации алгоритма машинного обучения является типичным примером таблицы сопряженности, в которой на диагонали указано число правильно классифицированных объектов, вне диагонали число ошибочно классифицированных объектов.

15. **Коллинеарность** — пары независимых переменных в регрессионном анализе высоко коррелируют, если их корреляции по модулю близки к единице.
16. **Критерий отношения дисперсий** — F-критерий Фишера-Снедекора, используется для проверки гипотез о равенстве дисперсий в популяции.
17. **Критерий хи-квадрат Пирсона** — используется в частотных данных, проверяет нулевую гипотезу, что нет связи между факторами, которые определяют таблицу сопряжённости. Также применяется для тестирования разницы в долях (пропорциях) данных.
18. **Лог-нормальное распределение** — вытянутое вправо распределение вероятности непрерывной случайной переменной, чей логарифм подчиняется нормальному распределению.
19. **Метод наименьших квадратов (МНК)** – метод оценки параметров в регрессионном анализе, основанный на минимизации суммы квадратов остатков.
20. **Межквартильный размах** – интервал между 25-й и 75-й процентилями; он содержит центральные 50% упорядоченных значений.
21. **Непараметрический критерий** – критерий проверки гипотез, который не делает предположений о распределении анализируемых данных. Иногда называется критерием, свободным от распределения.
22. **Несмещённая оценка** - для того чтобы оценка была несмещенной, требуется, чтобы в среднем оценка дала истинное значение неизвестного параметра. Формально оценка X является несмещенной оценкой параметра θ, если $E(X) = \theta$.
23. **Номограмма Альтмана** – диаграмма, которая устанавливает связь размера выборки, мощности статистического критерия, уровень значимости и стандартизированную разность.
24. **Перекрёстные исследования** – исследования, в которых каждый исследуемых пациент получает более одного вида лечения, одно за другим в случайном порядке.
25. **Пересечение множеств** — пересечение множеств A и B, обозначаемых A B, является множеством элементы, которые находятся как в A, так и в B.
26. **Протокол (Protocol)** - набор правил, определяющий все, что связано с работой сети.
27. **Пуассоновская регрессия** — в пуассоновской регрессии предполагается, что зависимая переменная распределена по закону Пуассона, где $\mu = E(Y | X)$ – среднее значение зависимой переменной Y при известных значениях независимых переменных X. В качестве функции связи обычно используется логарифм, также степенную и тождественную функцию.
28. **Размер выборки** — количество элементов в выборке. Размер выборки является важной величиной, при увеличении размера выборок точность

оценок увеличивается. Однако мы не можем увеличивать размер выборки до бесконечности, так это связано с временными и финансовыми затратами.

29. **Разнообразие Больших Данных (Big Data Variety)** — относится к типу и характеру данных. Это помогает людям, которые анализируют его, эффективно использовать полученную информацию.
30. **Сезонная вариация** – значение интересующей нас переменной систематически изменяются согласно времени года.
31. **Сервер (Server)** - компьютер (или программа), которая оказывает некоторые услуги клиентам - другим компьютерам (программам).
32. **Сериальная корреляция** – корреляция между наблюдениями во временных сериях и наблюдениями, отделёнными между собой фиксированным временным интервалом.
33. **Событие** — подмножество пространства выборки. Например, пространство для эксперимента, в котором дважды бросается монета, определяется {OO, OP, PO, PP} и A = {OP, OO}, тогда A событие, в котором Орёл встречается в первую очередь.
34. **Статистический критерий Вальда** — применяется в логистической регрессии для проверки вклада отдельного коэффициента регрессии.
35. **Форест-график** – диаграмма, применяемая в метаанализе и показывающая оценённый эффект в каждом исследовании и их среднее с доверительными интервалами.
36. **Хи-квадрат критерий** — используется для проверки гипотезы об отсутствии между факторами в таблице сопряжённости. Также используется для проверки различий между пропорциями (долями) в данных, проверки однородности.
37. **Цензурированные (неполные) данные** — используются в анализе выживаемости, поскольку имеется неполная информация об исходе лечения. Также используются в оценке надежности технических систем.
38. **«Ящик с усами»** — диаграмма, построенная из набора числовых данных, в центре которой находится медиана, по сторонам ящика – квартили (максимальные и минимальные значения).