

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**
**Федеральное государственное бюджетное
образовательное учреждение высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

СТАТИСТИЧЕСКИЕ ОСНОВЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ

**Англо-русский, русско-английский
словарь-справочник**

STATISTICAL FUNDAMENTALS OF BIG DATA ANALYSIS

**English-Russian, Russian-English
reference dictionary**

ПЕНЗА 2026

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Пензенский государственный университет» (ПГУ)

Статистические основы анализа больших данных

Англо-русский, русско-английский
словарь-справочник

Составитель

И. М. Мойко

Statistical fundamentals of big data analysis

English-Russian, Russian-English
reference dictionary

Compiler

I. M. Moiko

Пенза

Издательство ПГУ

2026

УДК 519.2
С78

Р е ц е н з е н т

кандидат физико-математических наук, доцент, доцент кафедры
«Математика и суперкомпьютерное моделирование»
Пензенского государственного университета
М. Ю. Медведик

Статистические основы анализа больших данных :
С78 англо-русский, русско-английский словарь-справочник /
сост. И. М. Мойко = Statistical fundamentals of big data analy-
sis : English-Russian, Russian-English reference dictionary /
comp. by I. M. Moiko. – Пенза : Изд-во ПГУ, 2026. – 144 с.

Содержатся ключевые термины и понятия, используемые в теории вероятностей и математической статистике в контексте анализа больших данных этой области, с переводом на английский язык. Также приведены пояснения и определения терминов, способствующие глубокому пониманию материала и освоению профессиональной лексики.

Издание подготовлено на кафедре «Высшая и прикладная математика» ПГУ и предназначено для англоязычных обучающихся по направлению подготовки 09.03.01 «Информатика и вычислительная техника», осваивающих дисциплины, связанные с анализом данных и математической статистикой, а также для преподавателей, работающих с англоязычными студентами.

This edition contains key terms and concepts used in probability theory and mathematical statistics in the context of big data analysis, with translations into Russian. The dictionary also includes explanations and definitions of terms, facilitating a deep understanding of the material and the acquisition of professional vocabulary.

The publication was prepared at the Department of Higher and Applied Mathematics of PSU and is intended for students in the field of study 09.03.01 “Informatics and Computer Engineering” taking courses related to data analysis and mathematical statistics, as well as for teachers working with English-speaking students.

УДК 519.2

© Пензенский государственный
университет, 2026

Contents

Introduction	4
Введение	5
SECTION 1. FUNDAMENTALS OF PROBABILITY THEORY AND STATISTICAL THINKING	6
1.1. Introduction to Big Data and the Role of Statistics in Artificial Intelligence	6
1.2. Probability Theory for Data Science	9
SECTION 2. EXPLORATORY DATA ANALYSIS AND DESCRIPTIVE STATISTICS	14
2.1. Fundamentals of Data Description	14
2.2. Statistical Estimation and Hypothesis Testing	21
SECTION 3. STATISTICAL MODELS FOR SUPERVISED LEARNING	29
3.1. Regression from a Statistical Point of View	29
3.2. Probabilistic Classification and Evaluation of Models	38
SECTION 4. ADVANCED TOPICS: UNSUPERVISED TECHNIQUES, SCALABILITY AND AI INTEGRATION	48
4.1. Statistical Approaches to Unsupervised Learning	48
4.2. Statistics in Modern AI Systems and New Trends	55
Index of Terms	68
РАЗДЕЛ 1. ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И СТАТИСТИЧЕСКОГО МЫШЛЕНИЯ	79
1.1. Введение в большие данные и роль статистики в искусственном интеллекте	79
1.2. Теория вероятностей для науки о данных	81
РАЗДЕЛ 2. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ И ОПИСАТЕЛЬНАЯ СТАТИСТИКА	87
2.1. Принципы описания данных	87
2.2. Статистическое оценивание и проверка гипотез	93
РАЗДЕЛ 3. СТАТИСТИЧЕСКИЕ МОДЕЛИ ДЛЯ ОБУЧЕНИЯ С УЧИТЕЛЕМ	98
3.1. Регрессия со статистической точки зрения	98
3.2. Вероятностная классификация и оценка моделей	105
РАЗДЕЛ 4. ПРОДВИНУТЫЕ ТЕМЫ: МЕТОДЫ БЕЗ УЧИТЕЛЯ, МАСШТАБИРУЕМОСТЬ И ИНТЕГРАЦИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА	114
4.1. Статистические подходы к обучению без учителя	114
4.2. Статистика в современных системах искусственного интеллекта и новые тенденции	120
Указатель терминов	130
Список литературы	142

Introduction

When studying mathematical disciplines, it is extremely important to adhere to strict mathematical definitions and terminology, as incorrect interpretation can lead to serious errors. Therefore, the glossaries presented to the reader serve several important functions:

1. **Provide clarity.** The glossary allows readers to learn the meanings of unfamiliar terms or terms used in the course. This makes it easier for readers to understand the content of the text.

2. **Prevent confusion.** If the text uses terms that have different meanings in different contexts, the glossary provides definitions that eliminate the possibility of misinterpretation.

3. **Ensure terminological consistency** for students both during classroom work and independent study.

This publication was prepared by the Department of Higher and Applied Mathematics of PSU and is intended for students majoring in 09.03.01 Informatics and Computer Engineering, who study the discipline “Statistical Fundamentals of Big Data Analysis” in English as a medium of instruction.

Введение

При изучении математических дисциплин крайне важно придерживаться строгих математических определений и терминологии, поскольку неправильная интерпретация может привести к серьезным ошибкам. В связи с этим представленные читателю глоссарии выполняют несколько важных функций:

1. Обеспечение ясности. Глоссарий позволяет читателям узнать значения незнакомых терминов или терминов, используемых в курсе. Это облегчает понимание содержания текста.

2. Предотвращение путаницы. Если в тексте используются термины, имеющие разные значения в разных контекстах, глоссарий предоставляет определения, исключающие возможность неправильной интерпретации.

3. Обеспечение терминологической согласованности для студентов как во время аудиторных занятий, так и при самостоятельной работе.

Данное издание подготовлено кафедрой высшей и прикладной математики ПГУ и предназначено для студентов направления подготовки 09.03.01 «Информатика и вычислительная техника», изучающих дисциплину «Статистические основы анализа больших данных» на английском языке.

SECTION 1

FUNDAMENTALS OF PROBABILITY THEORY AND STATISTICAL THINKING

1.1. Introduction to Big Data and the Role of Statistics in Artificial Intelligence

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
BIG DATA	Наборы данных чрезвычайно большого объема и сложности, которые требуют специальных методов обработки и анализа. В контексте искусственного интеллекта большие данные представляют собой фундаментальный ресурс для обучения моделей машинного обучения, позволяя алгоритмам выявлять сложные паттерны и зависимости	Система распознавания лиц «Безопасный город» обрабатывает миллионы изображений с камер видеонаблюдения для обучения нейронной сети, которая автоматически идентифицирует людей в потоке видео
VOLUME	Характеристика, описывающая масштаб данных. В задачах ИИ* объем данных напрямую влияет на качество моделей: чем больше размеченных данных, тем точнее алгоритм может обучаться, однако это также требует больших вычислительных ресурсов и памяти	Для обучения модели перевода Яндекс Переводчика используются миллиарды пар предложений на разных языках, что требует петабайтов хранилища данных
VELOCITY	Характеристика, описывающая скорость генерации и обработки данных. В системах реального времени, таких как автономные автомобили или алгоритмы	Система рекомендаций Яндекс Музыки анализирует действия пользователя (прослушивания, пропуски, лайки) в реальном времени и мгновенно адаптирует

* – искусственный интеллект.

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	высокочастотной торговли, данные должны обрабатываться за миллисекунды для принятия мгновенных решений	плейлист под текущие предпочтения
VARIETY	Характеристика, описывающая различные форматы и типы данных. Современные системы ИИ работают с мультимодальными данными: текстом, изображениями, видео, аудио, числовыми показателями, геопространственными данными и временными рядами	Мультимодальная модель YandexGPT анализирует одновременно текстовые запросы, изображения и голосовые команды для предоставления комплексных ответов пользователю
ARTIFICIAL INTELLIGENCE	Область компьютерных наук, занимающаяся созданием систем, способных выполнять задачи, требующие человеческого интеллекта: распознавание образов, понимание естественного языка, принятие решений, обучение на опыте и решение сложных проблем	Голосовой помощник «Алиса» использует искусственный интеллект для распознавания речи, понимания намерений пользователя, поиска информации и ведения диалога на естественном языке
MACHINE LEARNING	Подраздел искусственного интеллекта, использующий статистические методы для автоматического улучшения производительности на основе данных. Алгоритмы машинного обучения строят математические модели на основе обучающих данных для выполнения прогнозов или принятия решений без явного программирования	Алгоритм машинного обучения Сбербанка анализирует историю кредитных заявок и платежей клиентов для прогнозирования вероятности дефолта нового заемщика
DATA SCIENCE	Междисциплинарная область, использующая научные методы, процессы, алгоритмы и системы для	Специалист по данным в Wildberries анализирует поведение покупателей, сезонность спроса и внешние

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	извлечения знаний и инсайтов из структурированных и неструктурированных данных. Data Science объединяет статистику, машинное обучение, анализ данных и предметную экспертизу	факторы для оптимизации складских запасов и прогнозирования продаж
STATISTICAL LEARNING	Набор инструментов для моделирования и понимания сложных наборов данных с помощью статистических методов. Включает методы регрессии, классификации, кластеризации и снижения размерности для выявления закономерностей в данных	Метод статистического обучения (линейная регрессия) используется в сервисе «Циан» для прогнозирования стоимости квартир на основе их характеристик: площади, этажа, района, года постройки
DATA MINING	Процесс обнаружения скрытых паттернов, аномалий, корреляций и знаний в больших объемах данных с использованием методов на стыке статистики, машинного обучения и анализа баз данных	Сеть магазинов «Пятерочка» применяет Data Mining для анализа чеков покупателей и обнаружения ассоциативных правил («Покупатели, которые берут пиво, часто покупают чипсы»), что оптимизирует размещение товаров
PREDICTIVE ANALYTICS	Использование статистических алгоритмов и методов машинного обучения для прогнозирования будущих событий на основе исторических данных и выявления вероятностных тенденций	Предиктивная аналитика в авиакомпании «Аэрофлот» прогнозирует спрос на билеты по разным направлениям, что позволяет динамически изменять цены и оптимизировать загрузку рейсов

1.2. Probability Theory for Data Science

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
RANDOM VARIABLE	Величина, которая в результате эксперимента (наблюдения) принимает одно из возможных значений, заранее неизвестное и зависящее от случайных причин. В машинном обучении случайные величины моделируют неопределенность данных и прогнозов	Количество кликов пользователя на рекламный баннер за день – случайная величина, используемая для оптимизации рекламных кампаний в Яндекс Директе
DISCRETE RANDOM VARIABLE	Случайная величина, способная принимать конечное или счетное множество значений. Дискретные величины характерны для задач классификации, подсчета событий и категориальных данных	Количество ошибок в коде программы перед успешной компиляцией – дискретная случайная величина, которую можно использовать для оценки качества разработки
CONTINUOUS RANDOM VARIABLE	Случайная величина, которая может принимать любые значения из некоторого интервала (конечного или бесконечного). Непрерывные величины применяются для моделирования физических измерений, времени, вероятностей	Время отклика сервера (в миллисекундах) – непрерывная случайная величина, критически важная для мониторинга производительности веб-сервисов
PROBABILITY DISTRIBUTION	Функция или правило, описывающее вероятность принятия случайной величиной различных возможных значений. Распределения вероятностей являются фундаментом статистического моделирования в ИИ	Распределение времени доставки заказов Яндекс Еды описывается логнормальным распределением: большинство доставок занимает 30–45 мин, но возможны и более длительные сроки
PROBABILITY MASS FUNCTION, PMF	Функция, задающая распределение вероятностей для дискретной случайной величины. PMF присваивает каждому возможному значению вероятность его появления, причем сумма всех вероятностей равна 1	Функция вероятности для количества обращений в службу поддержки за 1 ч: $P(0) = 0,1$; $P(1) = 0,2$; $P(2) = 0,3$; $P(3) = 0,25$; $P(4) = 0,15$

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
PROBABILITY DENSITY FUNCTION, PDF	Функция, описывающая плотность вероятности для непрерывной случайной величины. Площадь под кривой PDF на интервале $[a, b]$ равна вероятности того, что случайная величина попадет в этот интервал	Плотность распределения времени жизни батареи смартфона описывает, что наиболее вероятно время работы 8–10 ч, а значения менее 4 или более 14 ч маловероятны
CUMULATIVE DISTRIBUTION FUNCTION, CDF	Функция, определяющая вероятность того, что случайная величина примет значение, меньшее заданного или равное ему. CDF используется для вычисления процентилей и квантилей в анализе данных	Функция распределения времени загрузки страницы показывает, что 90 % пользователей видят страницу полностью за 2 с или быстрее (метрика Page Load Time)
EXPECTED VALUE	Среднее взвешенное значение случайной величины, где весами служат вероятности. Математическое ожидание характеризует «центр» распределения и используется как прогноз в задачах регрессии	Математическое ожидание количества продаж товара в день равно 15 единицам – это значение используется для оптимизации складских запасов
VARIANCE	Мера разброса случайной величины вокруг ее математического ожидания. Дисперсия показывает степень неопределенности или риска: высокая дисперсия означает большую вариативность результатов	Дисперсия доходности инвестиционного портфеля показывает риск: 2 портфеля с одинаковой ожидаемой доходностью 10 % могут иметь разную дисперсию – 2 % (стабильный) или 15 % (рискованный)
STANDARD DEVIATION	Квадратный корень из дисперсии, измеряемый в тех же единицах, что и сама случайная величина. Стандартное отклонение интерпретируется как типичное отклонение от среднего и используется для оценки точности моделей	Стандартное отклонение времени доставки еды составляет 8 мин: если среднее время 35 мин, то большинство доставок занимает от 27 до 43 мин
CONDITIONAL PROBABILITY	Вероятность события при условии, что другое событие уже произошло. Условные вероятности лежат в основе	Вероятность того, что пользователь купит товар (событие A), при условии, что он добавил его в корзину

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	байесовских сетей, наивного байесовского классификатора и многих алгоритмов ИИ	(событие B), равна 0,65 – это используется для прогнозирования конверсии в интернет-магазинах
BAYES THEOREM	Фундаментальная теорема теории вероятностей, количественно определяющая вероятность события на основе предварительных знаний об условиях, которые могут быть связаны с событием. Теорема Байеса является основой байесовского машинного обучения и фильтрации спама	Система обнаружения мошеннических транзакций использует теорему Байеса: если транзакция совершена в необычном месте и на крупную сумму, вероятность мошенничества возрастает с 0,1 % (априорная) до 78 % (апостериорная)
INDEPENDENCE	Свойство событий, при котором наступление одного события не влияет на вероятность другого. Предположение о независимости упрощает вычисления в вероятностных моделях, но в реальных данных часто наблюдается зависимость	Результаты бросания игровой кости в разные моменты времени независимы – это свойство используется при генерации случайных чисел для инициализации весов нейронных сетей
RANDOM EXPERIMENT	Процесс, результат которого непредсказуем, но множество всех возможных результатов известно. В машинном обучении случайные эксперименты моделируют стохастичность данных и алгоритмов	A/B -тестирование 2 версий лендинга – случайный эксперимент, где результат (конверсия) зависит от случайного распределения пользователей между вариантами
SAMPLE SPACE	Множество всех возможных исходов случайного эксперимента. Пространство исходов определяет область значений, в которой работает вероятностная модель	Для задачи бинарной классификации (спам/не спам) пространство исходов состоит из 2 элементов: {спам, не спам}. Для распознавания цифр MNIST – {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
EVENT	Подмножество пространства элементарных исходов, обладающее определенным свойством. События используются	Событие «пользователь совершил покупку» включает следующие исходы: покупка на любую сумму, в любой категории

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	для формулировки гипотез в статистическом тестировании и A/B -экспериментах	товара, любым способом оплаты
JOINT PROBABILITY	Вероятность одновременного наступления 2 или более событий. Совместные распределения вероятностей моделируют взаимосвязи между признаками в многомерных данных	Совместная вероятность того, что пользователь одновременно кликнет на рекламу и купит товар, равна 0,03 – это используется для расчета ожидаемой прибыли от рекламной кампании
MARGINAL PROBABILITY	Вероятность события без учета других событий, полученная суммированием (или интегрированием) совместной вероятности по всем возможным значениям других переменных	Вероятность покупки товара (независимо от того, кликал ли пользователь на рекламу или пришел органически) равна 0,12 – это маргинальная вероятность конверсии
BERNOULLI DISTRIBUTION	Дискретное распределение с 2 возможными исходами: успех (1) с вероятностью p и неудача (0) с вероятностью $1 - p$. Распределение Бернулли моделирует бинарные события в классификации	Результат предсказания модели: кликнет ли пользователь на рекламу (1 с вероятностью $p = 0,25$) или нет (0 с вероятностью 0,75)
BINOMIAL DISTRIBUTION	Распределение числа успехов в последовательности n независимых испытаний Бернулли с одинаковой вероятностью успеха p . Биномиальное распределение используется для A/B -тестирования и оценки точности классификаторов	Вероятность того, что из 100 пользователей сайта ровно 15 совершат покупку при конверсии 12 %, описывается биномиальным распределением
NORMAL DISTRIBUTION	Непрерывное симметричное распределение относительно среднего значения, характеризующееся параметрами μ (среднее) и σ^2 (дисперсия). Нормальное распределение фундаментально для статистического вывода и многих алгоритмов машинного обучения	Рост пользователей мобильного приложения подчиняется нормальному распределению со средним (170 см) и стандартным (10 см) отклонениями

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
POISSON DISTRIBUTION	Распределение, описывающее число событий, происходящих за фиксированный интервал времени, при условии, что события происходят независимо с постоянной средней интенсивностью. Распределение Пуассона моделирует редкие события	Количество сбоев сервера в час (в среднем 0,5 сбоя) описывается распределением Пуассона – это используется для планирования ресурсов и SLA*
UNIFORM DISTRIBUTION	Распределение, при котором все исходы равновероятны. Равномерное распределение используется для инициализации весов нейронных сетей и генерации случайных чисел	При инициализации весов нейронной сети значения выбираются из равномерного распределения на интервале $[-0,1; 0,1]$
LAW OF LARGE NUMBERS	Теорема о том, что среднее значение выборки сходится к математическому ожиданию при увеличении размера выборки. Закон больших чисел обосновывает необходимость больших данных для обучения моделей ИИ	Точность предсказания модели стабилизируется при увеличении тестовой выборки: на 10 примерах точность может быть 70 %, на 1000 примерах – 84 %, на 10 000 примерах – 85,2 % (приближается к истинному значению)
CENTRAL LIMIT THEOREM, CLT	Теорема о том, что сумма большого числа независимых случайных величин имеет приблизительно нормальное распределение независимо от распределений отдельных слагаемых. Центральная предельная теорема обосновывает использование нормального распределения в статистических тестах и доверительных интервалах	Средняя оценка приложения в App Store (на основе тысяч отзывов) имеет нормальное распределение, даже если отдельные оценки распределены неравномерно, – это позволяет строить доверительные интервалы для среднего рейтинга

* Service Level Agreement.

SECTION 2

EXPLORATORY DATA ANALYSIS AND DESCRIPTIVE STATISTICS

2.1. Fundamentals of Data Description

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
EXPLORATORY DATA ANALYSIS, EDA	Подход к анализу наборов данных для обобщения их основных характеристик, часто с использованием визуальных методов. EDA позволяет понять структуру данных, выявить аномалии, проверить предположения и определить важные переменные перед построением моделей машинного обучения	В практике data scientist'a разведочный анализ помогает выявить выбросы и аномалии перед обучением моделей: например, при анализе данных о транзакциях можно обнаружить ошибочные значения сумм покупок (отрицательные числа или нереалистично большие величины), которые нужно исправить или удалить
DESCRIPTIVE STATISTICS	Методы суммирования и описания основных характеристик набора данных с помощью числовых показателей и графиков. Описательная статистика включает расчет мер центральной тенденции (среднее, медиану, моду), мер разброса (дисперсию, стандартное отклонение, размах), а также визуализацию распределений	При анализе базы данных клиентов Интернет-магазина описательная статистика показывает следующее: средний чек составляет 2 500 руб., медианный чек – 1 800 руб., стандартное отклонение – 1 200 руб. Это говорит о значительном разбросе сумм покупок
MEASURES OF CENTRAL TENDENCY	Статистики, описывающие центр распределения данных или типичное значение	При анализе зарплат сотрудников IT-компании средняя зарплата может составлять 150 000 руб.,

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	в наборе данных. Основные меры: среднее арифметическое (<i>mean</i>), медиана (<i>median</i>) и мода (<i>mode</i>). Выбор меры зависит от типа данных и наличия выбросов	но из-за высоких зарплат топ-менеджеров медиана равна 95 000 руб., что лучше отражает типичный уровень дохода большинства сотрудников
MEAN	Сумма всех значений, поделенная на их количество. Среднее арифметическое чувствительно к выбросам и экстремальным значениям, что может искажать представление о типичном значении в данных с асимметричным распределением	Время отклика мобильного приложения за день: если зафиксированы значения 120, 130, 125, 140 мс и 1 сбой – 2500 мс, то среднее составит 603 мс, что не отражает типичную производительность системы
MEDIAN	Среднее значение упорядоченной выборки: если количество наблюдений нечетное, медиана – это центральное значение; если четное – среднее 2 центральных значений. Медиана устойчива к выбросам и лучше характеризует типичное значение в асимметричных распределениях	При анализе времени доставки еды курьерами за смену: из 100 доставок 95 заняли от 20 до 40 мин, а 5 доставок – более 90 мин из-за сложных заказов. Медианное время 28 мин лучше отражает типичную скорость доставки, чем среднее значение 35 мин
MODE	Наиболее часто встречающееся значение в наборе данных. Мода особенно полезна для категориальных данных и дискретных величин. В мультимодальных распределениях может быть несколько мод	В логах ошибок веб-сервиса за неделю наиболее частая ошибка (мода) – Timeout database connection (встречается 247 раз), что указывает на проблему с пулом соединений базы данных и помогает приоритезировать исправления

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
MEASURES OF DISPERSION	Статистики, описывающие вариабельность данных, т.е. насколько значения разбросаны относительно центра. Основные меры: размах (<i>range</i>), дисперсия (<i>variance</i>), стандартное отклонение (<i>standard deviation</i>), межквартильный размах (<i>IQR</i>)	Две команды разработки показывают одинаковую среднюю скорость выполнения задач – 8 <i>story points</i> в спринт. Но у 1-й команды стандартное отклонение – 1,5 <i>points</i> (стабильная работа), а у 2-й – 4 <i>points</i> (непредсказуемая производительность), что критично для планирования релизов
RANGE	Разность между максимальным и минимальным значениями в наборе данных. Размах дает простую оценку вариабельности, но крайне чувствителен к выбросам и не учитывает распределение значений внутри диапазона	Диапазон времени загрузки страниц сайта: минимальное время – 0,8 с, максимальное – 45 с (из-за редких сбоев). Размах 44,2 с создает ложное впечатление о нестабильности, тогда как 99 % запросов загружаются за 0,8–2,5 с
INTERQUARTILE RANGE, IQR	Разность между 3-м (Q_3 , 75-й перцентиль) и 1-м (Q_1 , 25-й перцентиль) квартилями. IQR показывает разброс средних 50 % данных и устойчив к выбросам, что делает его предпочтительной мерой разброса для асимметричных распределений	При анализе зарплат junior-разработчиков Python: $Q_1 = 60\ 000$ руб., $Q_3 = 95\ 000$ руб., $IQR = 35\ 000$ руб. Это означает, что средние 50 % специалистов получают зарплаты в диапазоне 60–95 тыс. руб., что дает реалистичное представление о рынке без влияния экстремально высоких и низких значений
QUANTILES	Точки деления упорядоченной выборки на равные части. Квантиль уровня p делит данные так, что $p \cdot 100$ % наблюдений лежат ниже этого значения. Частные случаи: квартили (деление на 4 части), перцентили (деление на 100 частей)	95-й перцентиль времени отклика API равен 450 мс: это означает, что 95 % запросов обрабатываются быстрее 450 мс. Этот показатель используется в SLA вместо среднего времени отклика, так как лучше отражает опыт большинства пользователей

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
PERCENTILES	Деление упорядоченных данных на 100 равных частей. n -й перцентиль показывает значение, ниже которого падает n % наблюдений. Перцентили широко используются для анализа производительности систем и установления пороговых значений	90-й перцентиль времени загрузки мобильного приложения составляет 2,3 с: 90 % пользователей видят полностью загруженный интерфейс быстрее этого значения, а оставшиеся 10 % сталкиваются с более медленной загрузкой из-за медленного Интернета или старых устройств
QUARTILES	Деление упорядоченных данных на 4 равные части. 1-й квартиль (Q_1) соответствует 25-му перцентилю, 2-й квартиль (Q_2) – медиане (50-й перцентиль), 3-й квартиль (Q_3) – 75-му перцентилю. Квартили описывают распределение данных и используются для выявления выбросов	Анализ оценок студентов за экзамен по машинному обучению: $Q_1 = 62$ балла; Q_2 (медиана) = 74 балла; $Q_3 = 85$ баллов. Анализ показывает, что 25 % студентов набрали менее 62 баллов, 50 % – менее 74 баллов и 75 % – менее 85 баллов. Это помогает оценить сложность экзамена
SKEWNESS	Мера асимметрии распределения данных относительно среднего значения. Положительная асимметрия (правосторонняя) указывает на длинный правый хвост распределения (большинство значений сконцентрировано слева), отрицательная – на левостороннюю асимметрию	Распределение доходов пользователей премиум-подписки стримингового сервиса имеет положительную асимметрию: большинство платит базовую цену 299 руб. в месяц, но небольшая группа корпоративных клиентов отдает десятки тысяч рублей, создавая длинный правый хвост распределения
KURTOSIS	Мера остроты пика распределения и тяжести хвостов по сравнению с нормальным распределением. Высокий эксцесс (положительный) указывает на тяжелые	Распределение ежедневной доходности акций технологической компании имеет высокий эксцесс: большинство дней показывают небольшую изменчивость (острый пик около 0 %), но периодически

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	хвосты и острый пик (больше выбросов), низкий эксцесс (отрицательный) – на плоское распределение с легкими хвостами	случаются резкие скачки цен на 10–15 % (тяжелые хвосты), что критично для оценки рисков
OUTLIER	Наблюдение, которое значительно отличается от других наблюдений в выборке. Выбросы могут указывать на ошибки в данных, редкие события или важные аномалии. В машинном обучении выбросы могут искажать модели и требуют специальной обработки	В системе мониторинга серверов зафиксирована температура процессора 125 °С при типичных значениях 45–70 °С. Этот выброс может указывать на сбой системы охлаждения, ошибку датчика или экстремальную нагрузку – требуется немедленная проверка для предотвращения повреждения оборудования
BOX PLOT	Графическое представление 5-числовой сводки данных: минимальное значение, 1-й квартиль (Q_1), медиана (Q_2), 3-й квартиль (Q_3) и максимальное значение. Ящик охватывает межквартильный размах (IQR), «усы» показывают диапазон данных, а точки за пределами «усов» – выбросы	При сравнении производительности 3 алгоритмов классификации боксплоты показывают следующее: алгоритм А имеет медианную точность 87 % с небольшим разбросом, алгоритм В – 89 %, но с большим разбросом, алгоритм С – 85 % с несколькими выбросами вниз. Это помогает выбрать стабильный алгоритм для продажи
HISTOGRAM	Графическое представление распределения числовых данных: диапазон значений делится на интервалы (бины), и для каждого интервала отображается частота (количество) попаданий. Гистограмма	Гистограмма времени сессий пользователей на сайте показывает бимодальное распределение: 1-й пик на 2–3 мин (быстрый просмотр и уход), 2-й пик на 15–20 мин (вовлеченные пользователи). Это помогает сегментировать

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	визуализирует форму распределения, модальность и асимметрию	аудиторию и оптимизировать контент
SCATTER PLOT	График для отображения взаимосвязи между 2 числовыми переменными: каждая точка представляет наблюдение с координатами (x, y) . Помогает выявить линейные или нелинейные зависимости, кластеры и выбросы в двумерных данных	Диаграмма рассеяния зависимости цены квартиры от площади показывает положительную корреляцию: с ростом площади цена увеличивается. Также видны выбросы – элитные квартиры с аномально высокой ценой за квадратный метр и кластеры по районам города
CORRELATION	Статистическая взаимосвязь между 2 переменными, показывающая, как изменение одной переменной связано с изменением другой. Корреляция не подразумевает причинно-следственную связь	Обнаружена сильная положительная корреляция между количеством просмотров видео на YouTube (запрещен в России) и числом лайков ($r = 0,78$). Однако это не означает, что лайки вызывают просмотры: оба показателя зависят от качества контента и популярности автора
PEARSON CORRELATION	Мера линейной корреляции между 2 переменными, принимающая значения от -1 до $+1$. Значение $+1$ указывает на идеальную положительную линейную зависимость; -1 – на идеальную отрицательную; 0 – на отсутствие линейной связи. Требуется нормальное распределение данных	Корреляция Пирсона между временем изучения курса по data science и итоговой оценкой составляет $r = 0,65$ (умеренная положительная корреляция): студенты, потратившие больше часов на материал, как правило, получают более высокие оценки
SPEARMAN CORRELATION	Мера монотонной (не обязательно линейной) связи между переменными, основанная на	При анализе удовлетворенности сотрудников (оценки от 1 до 5) и их производительности (количество

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	рангах значений, а не на самих значениях. Корреляция Спирмена используется для ненормально распределенных данных, порядковых шкал и при наличии выбросов	выполненных задач) корреляция Спирмена ($\rho = 0,42$) показывает умеренную положительную связь: более довольные сотрудники склонны работать продуктивнее, даже если зависимость нелинейна
COVARIANCE	Мера совместной изменчивости 2 случайных величин, показывающая направление их линейной связи. Положительная ковариация указывает на изменение переменных в одном направлении, отрицательная – в противоположных. В отличие от корреляции, ковариация зависит от масштаба данных	Ковариация между ценой акции компании и объемом торгов положительная: когда цена растет, обычно увеличивается и активность торгов. Однако абсолютное значение ковариации трудно интерпретировать без нормализации, поэтому для сравнения силы связи используют коэффициент корреляции
DATA VISUALIZATION	Графическое представление данных и информации с использованием визуальных элементов (графиков, диаграмм, карт, дашбордов). Визуализация помогает быстро выявлять паттерны, аномалии, тренды и взаимосвязи, которые трудно обнаружить при анализе сырых данных	Интерактивный дашборд в Tableau показывает ключевые метрики Интернет-магазина: динамику продаж, географию заказов, конверсию воронки, топ товаров. Менеджеры за несколько секунд оценивают ситуацию и принимают решения, вместо изучения таблиц с тысячами строк
MISSING DATA	Отсутствующие значения в наборе данных, которые могут возникнуть из-за ошибок сбора, отказа пользователей отвечать на вопросы, сбоя оборудования.	В опросе клиентов о доходе 30 % респондентов не указали сумму. Простое удаление этих строк сместит выборку в сторону более открытых клиентов. Лучше использовать множественную импутацию

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	Пропущенные данные приводят к смещенным оценкам и снижению качества моделей, если не обработаны корректно	(<i>Multiple Imputation</i>) или добавить индикатор пропуска как отдельный признак
DATA CLEANING	Процесс обнаружения и исправления ошибок в данных: обработка пропущенных значений, удаление дубликатов, исправление некорректных форматов, устранение выбросов, стандартизация категорий. Очистка данных занимает 60–80 % времени в проектах data science	При подготовке данных о транзакциях для ML-модели очистка включает удаление тестовых транзакций с суммой 0 руб., исправление дат в неверном формате, объединение дублирующихся категорий («Москва» и «москва»), замену выбросов (покупки на 10 млн руб.) на максимальное реалистичное значение

2.2. Statistical Estimation and Hypothesis Testing

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
STATISTICAL INFERENCE	Процесс использования данных выборки для получения выводов о генеральной совокупности: оценивания параметров, проверки гипотез, построения доверительных интервалов. Статистический вывод позволяет обобщать и распространять результаты эксперимента на всю популяцию с учетом неопределенности	На основе опроса 1500 пользователей мобильного приложения (выборка) делаем вывод о всей базе из 2 млн пользователей (генеральная совокупность): 68 ± 3 % довольны новым интерфейсом с доверительной вероятностью 95 %
POINT ESTIMATION	Оценка параметра генеральной совокупности одним числом на основе	Среднее время доставки курьеров за неделю составило 32 мин – это

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	выборки. Примеры точечных оценок: выборочное среднее (оценка математического ожидания), выборочная дисперсия (оценка дисперсии генеральной совокупности), выборочная доля (оценка вероятности события)	точечная оценка среднего времени доставки для всей генеральной совокупности заказов. Однако точечная оценка не показывает неопределенность: истинное среднее может отличаться
INTERVAL ESTIMATION	Оценка параметра в виде интервала значений (доверительного интервала), который с заданной вероятностью содержит истинное значение параметра. Интервальная оценка информативнее точечной, так как показывает неопределенность оценки	Доверительный интервал для конверсии лендинга: $12,5 \pm 1,8 \%$ (при доверительной вероятности 95 %). Это означает, что истинная конверсия генеральной совокупности с вероятностью 95 % лежит в диапазоне от 10,7 до 14,3 %, а не равна точно 12,5 %
CONFIDENCE INTERVAL	Интервал значений, который с заданной доверительной вероятностью (<i>confidence level</i>) содержит истинное значение оцениваемого параметра генеральной совокупности. Ширина интервала зависит от объема выборки, вариабельности данных и выбранного уровня доверия	95%-й доверительный интервал для средней зарплаты Python-разработчиков в Москве: $145\,000 \pm 12\,000$ руб. Интерпретация: если многократно повторять выборку и строить интервалы, то 95 % таких интервалов будут содержать истинное среднее значение зарплаты
CONFIDENCE LEVEL	Вероятность того, что доверительный интервал содержит истинное значение параметра. Распространенные уровни доверия: 90, 95, 99 %. Более высокий уровень доверия дает более широкий интервал (большую неопределенность), но большую надежность вывода	При A/B-тестировании новой функции используем 95%-й уровень доверия: если доверительный интервал разницы конверсий не включает ноль, делаем вывод о статистической значимости эффекта. Выбор 95 % (а не 99 %) – компромисс между надежностью и чувствительностью теста

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
HYPOTHESIS TESTING	Процедура проверки утверждений о параметрах генеральной совокупности на основе выборочных данных. Проверка гипотез включает формулировку нулевой и альтернативной гипотез, выбор уровня значимости, расчет тестовой статистики и решение об отклонении или непринятии нулевой гипотезы	Проверка гипотезы об эффективности нового алгоритма рекомендаций: нулевая гипотеза H_0 – средний чек пользователей не изменился, альтернативная H_1 – средний чек увеличился. После сбора данных и расчета ($p\text{-value} = 0,023$) делаем вывод о статистической значимости улучшения
NULL HYPOTHESIS, H_0	Гипотеза, которую проверяют и пытаются отвергнуть. Нулевая гипотеза обычно утверждает отсутствие эффекта, различий или связи между переменными («статус-кво»). В машинном обучении H_0 часто утверждает, что новая модель не лучше базовой	При сравнении 2 моделей классификации нулевая гипотеза такова: точность новой модели равна точности текущей продакшен-модели (разница = 0). Альтернативная гипотеза: новая модель точнее. Если $p\text{-value} < 0,05$, отвергаем нулевую гипотезу и внедряем новую модель
ALTERNATIVE HYPOTHESIS, H_1	Гипотеза, противоречащая нулевой. Альтернативная гипотеза утверждает наличие эффекта, различий или связи. Может быть односторонней (параметр больше или меньше) или двусторонней (параметр не равен)	В тестировании новой кнопки на сайте альтернативная гипотеза такова: конверсия новой версии выше текущей (односторонний тест). Если данные показывают статистически значимое увеличение с $p\text{-value} = 0,01$, принимаем H_1 и запускаем новую версию для всех пользователей
SIGNIFICANCE LEVEL, α	Вероятность отвергнуть нулевую гипотезу, когда она верна (ошибка I рода). Уровень значимости – это порог для	При запуске A/B-теста устанавливаем $\alpha = 0,05$. Если $p\text{-value} = 0,03 (< 0,05)$, отвергаем нулевую гипотезу об отсутствии

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	<i>p-value</i> , ниже которого результат считается статистически значимым. Стандартный уровень $\alpha = 0,05$ (5 %) означает готовность принять 5 % ложноположительных результатов	разницы и делаем вывод о значимости эффекта. Риск: с вероятностью 5 % мы можем сделать ошибочное заключение об эффекте, когда его нет
P-VALUE	Вероятность получить наблюдаемые или более экстремальные результаты при условии, что нулевая гипотеза верна. Малое <i>p-value</i> (меньше уровня значимости α) указывает на несовместимость данных с H_0 и становится основанием для ее отклонения. <i>p-value</i> НЕ является вероятностью того, что H_0 верна	После сравнения 2 алгоритмов получили <i>p-value</i> = 0,008. Это означает следующее: если бы алгоритмы были одинаково эффективны (H_0 верна), вероятность получить наблюдаемую разницу в точности или более extreme составила бы всего 0,8 %. Делаем вывод о статистической значимости различия
TYPE I ERROR	Отклонение верной нулевой гипотезы (<i>false positive</i>). Вероятность ошибки I рода равна уровню значимости α . В бизнесе ошибка I рода – это внедрение изменения, которое на самом деле не работает, что ведет к напрасным затратам	Запуск новой функции приложения показал статистически значимое увеличение вовлеченности ($p < 0,05$), и мы внедрили ее для всех пользователей. Через месяц метрики вернулись к исходным значениям: первоначальный эффект оказался ложноположительным результатом (ошибка I рода), стоимость разработки потрачена зря
TYPE II ERROR	Принятие (неотклонение) неверной нулевой гипотезы (<i>false negative</i>). Ошибка II рода означает пропуск реального эффекта. Вероятность ошибки II рода обозна-	Новый алгоритм детекции мошеннических транзакций действительно на 15 % точнее текущего, но из-за малой выборки (всего 200 транзакций) тест не выявил статистической значимости

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	чается β . Мощность теста ($Power$) = $1 - \beta$	$p = 0,12$. Мы отказались от внедрения эффективного алгоритма – это ошибка II рода
POWER OF TEST	Вероятность правильного отклонения ложной нулевой гипотезы $1 - \beta$. Мощность показывает способность теста обнаружить эффект, если он действительно существует. Для надежных результатов мощность должна быть не менее 0,8 (80 %)	Перед запуском A/B-теста рассчитываем необходимый размер выборки для достижения мощности 0,8 при ожидаемом эффекте 5 % и $\alpha = 0,05$. Получаем следующее: нужно минимум 2000 пользователей в каждой группе. Меньшая выборка рискует пропустить реальный эффект (низкая мощность)
TEST STATISTIC	Статистика, рассчитываемая по выборке для проверки гипотезы (t -статистика, z -статистика, χ^2 , F -статистика). Тестовая статистика сравнивается с критическим значением или используется для расчета p -value. Распределение тестовой статистики известно при верности H_0	При сравнении средних 2 групп рассчитали, что t -статистика = 2,47. По таблице распределения Стьюдента находим критическое значение для $\alpha = 0,05$ и $df = 58$: $t_{crit} = 2,00$. Так как $ 2,47 > 2,00$, отвергаем нулевую гипотезу о равенстве средних
CRITICAL REGION	Множество значений тестовой статистики, при которых нулевая гипотеза отвергается. Критическая область определяется уровнем значимости α и альтернативной гипотезой (односторонняя или двусторонняя). Если тестовая статистика попадает в критическую область, результат статистически значим	Для двустороннего z -теста при $\alpha = 0,05$ критическая область: $z < -1,96$ или $z > 1,96$. Рассчитанная z -статистика, равная 2,31, попадает в критическую область ($2,31 > 1,96$), поэтому отвергаем H_0 на уровне значимости 5 %
T-TEST	Статистический тест для сравнения средних значений, используя	Сравниваем среднее время отклика API до и после оптимизации:

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	щий t -распределение Стьюдента. Применяется, когда дисперсия генеральной совокупности неизвестна и оценивается по выборке (особенно важно для малых выборок $n < 30$)	выборка из 25 измерений показала t -статистику, равную 3,12, при $df = 24$. $p\text{-value} = 0,004 < 0,05$, следовательно, делаем вывод о статистически значимом улучшении производительности
Z-TEST	Тест, основанный на нормальном распределении (z -распределении). Применяется для больших выборок ($n \geq 30$) или в случае, когда известна дисперсия генеральной совокупности. Для пропорций и средних при больших n z -тест и t -тест дают схожие результаты	Тестируем гипотезу о конверсии лендинга: из 5000 посетителей 625 совершили целевое действие (12,5 %). При известной дисперсии и большом n используем z -тест. z -статистика = 1,89, $p\text{-value} = 0,059$. Следовательно, разница с целевой конверсией 11 % незначима на уровне 5 %
CHI-SQUARE TEST	Тест для проверки независимости 2 категориальных переменных или соответствия распределения теоретическому. χ^2 -тест сравнивает наблюдаемые частоты с ожидаемыми частотами при верности нулевой гипотезы	Проверяем зависимость между полом пользователя и выбором тарифа (базовый/премиум). Таблица сопряженности 2×2 дает следующее: $\chi^2 = 8,73$, $df = 1$, $p\text{-value} = 0,003$. Отвергаем гипотезу о независимости: мужчины и женщины статистически значимо различаются в выборе тарифа
ANOVA	Метод сравнения средних значений 3 и более групп одновременно. Проверяет нулевую гипотезу о равенстве всех групповых средних. Если ANOVA показывает значимость, проводятся пост-хок-тесты	Сравниваем эффективность 3 маркетинговых каналов (e-mail, соцсети, контекст) по средней стоимости привлечения клиента. One-way ANOVA: $F(2,87) = 5,34$, $p = 0,006$. Каналы статистически значимо различаются.

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	(Tukey, Bonferroni) для попарных сравнений	Пост-хок-тест Tukey показывает следующее: контекст дороже e-mail $p = 0,004$, остальные пары не различаются
F-TEST	Тест, использующий F -распределение для сравнения дисперсий 2 выборок или в ANOVA для сравнения нескольких средних. F -статистика – отношение 2 дисперсий. При сравнении дисперсий F близко к 1 указывает на их равенство	В ANOVA для сравнения 4 моделей машинного обучения: F -статистика = 4,67; $df_between = 3$; $df_within = 96$; $p = 0,005$. Модели статистически значимо различаются по точности. Наибольшую среднюю точность показывает градиентный бустинг (92,3 %)
DEGREES OF FREEDOM, DF	Число независимых значений в расчете статистики, равное объему выборки, минус число оцененных параметров. Степени свободы влияют на форму t -распределения, χ^2 -распределения и F -распределения. При больших df t -распределение приближается к нормальному	Для t -теста 2 независимых выборок размером $n_1 = 30$ и $n_2 = 35$ степени свободы $df = n_1 + n_2 - 2 = 63$. По таблице t -распределения для $df = 63$ и $\alpha = 0,05$ (двусторонний) находим критическое значение $t_{crit} = 1,998$
EFFECT SIZE	Количественная мера силы изучаемого явления или величины различий, независимая от объема выборки. Популярные меры: Cohen's d (для разности средних), η^2 (<i>eta-squared</i> для ANOVA), <i>odds ratio</i> (для бинарных исходов). Размер эффекта показывает практическую значимость, а не только статистическую	Сравнили 2 модели: новая модель статистически значимо точнее ($p = 0,001$), но Cohen's $d = 0,12$ (очень малый эффект). Разница в точности 0,3 % статистически значима из-за большой выборки, но практически не существенна для бизнеса – не стоит тратить ресурсы на замену модели

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
SAMPLE SIZE	Количество наблюдений в выборке. Большой размер выборки обеспечивает более точные оценки, узкие доверительные интервалы и высокую мощность тестов. Однако после определенного предела прирост точности становится незначительным (закон убывающей отдачи)	Для обнаружения увеличения конверсии на 2 % (с 10 до 12 %) при мощности 0,8 и $\alpha = 0,05$ нужен размер выборки 3 853 наблюдения в каждой группе. Меньшая выборка (например, 500) не обнаружит эффект, даже если он реален (низкая мощность)

SECTION 3

STATISTICAL MODELS FOR SUPERVISED LEARNING

3.1. Regression from a Statistical Point of View

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
SUPERVISED LEARNING	Тип машинного обучения, при котором алгоритм обучается на размеченных данных, т.е. данных, для которых известны правильные ответы (целевые переменные). Модель учится находить взаимосвязь между входными признаками и выходным значением, чтобы делать прогнозы для новых, ранее не виденных данных	Система кредитного скоринга банка обучается на исторических данных о заемщиках: для каждого клиента известны признаки (возраст, доход, кредитная история) и целевая переменная (вернул кредит или нет). После обучения модель предсказывает вероятность дефолта для новых заявителей на кредит
REGRESSION	Метод моделирования взаимосвязи между зависимой (целевой) переменной и одной или несколькими независимыми переменными (признаками). Регрессионный анализ используется для прогнозирования непрерывных числовых значений, а не категориальных меток	Сервис «Яндекс Недвижимость» использует регрессию для предсказания стоимости квартир: на основе признаков (площадь, этаж, район, год постройки) модель предсказывает рыночную цену в рублях – непрерывную числовую величину
DEPENDENT VARIABLE	Переменная, которую модель предсказывает или объясняет; также называется целевой переменной (<i>target</i>) или откликом (<i>response</i>). Зависимая переменная	В задаче прогнозирования спроса на товары в магазине Wildberries зависимой переменной выступает количество продаж конкретного товара за день – именно это

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	«зависит» от значений независимых переменных (признаков)	значение мы хотим предсказать на основе признаков (день недели, цена, наличие акций, сезонность)
INDEPENDENT VARIABLE	Переменная, используемая для предсказания зависимой переменной; также называется признаком (<i>feature</i>), предиктором или объясняющей переменной. Независимые переменные считаются входными данными модели	При предсказании времени доставки еды сервисом «Яндекс Еда» независимыми переменными выступают расстояние от ресторана до клиента, время суток, день недели, погода, количество заказов у курьера, средняя скорость движения в районе – все эти признаки влияют на целевую переменную (время доставки)
FEATURE	Отдельная измеримая характеристика наблюдаемого явления; отдельный столбец в наборе данных. Признаки могут быть числовыми (возраст, доход), категориальными (пол, город) или бинарными (наличие скидки: да/нет)	В системе распознавания спама e-mail признаки включают количество слов в письме, наличие определенных ключевых слов («выигрыш», «бесплатно»), количество ссылок, домен отправителя, время отправки, наличие вложений – каждый признак помогает модели отличить спам от легитимных писем
TARGET VARIABLE	Переменная, которую модель пытается предсказать; синоним зависимой переменной. В задачах классификации целевая переменная принимает категориальные значения (классы), в задачах регрессии – непрерывные числовые значения	В задаче обнаружения мошеннических транзакций Сбербанка целевая переменная принимает 2 значения: 1 (мошенничество) или 0 (легитимная операция). Модель анализирует признаки транзакции (сумма, место, время, тип операции) и предсказывает вероятность принадлежности к классу мошенничества

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
LINEAR REGRESSION	Регрессионный анализ, моделирующий линейную взаимосвязь между зависимой переменной и одной или несколькими независимыми переменными. Линейная регрессия предполагает, что изменение зависимой переменной пропорционально изменению предикторов	Модель предсказывает расход бензина автомобиля на основе пройденного расстояния: если автомобиль тратит 8 л на 100 км, то линейная зависимость имеет вид $y = 0,08x$, где x – расстояние, км; y – расход топлива, л
SIMPLE LINEAR REGRESSION	Регрессия с одной независимой переменной (одним признаком). Простая линейная регрессия описывается уравнением $y = \beta_0 + \beta_1 x + \varepsilon$, где β_0 – свободный член (<i>intercept</i>); β_1 – коэффициент наклона; ε – случайная ошибка	Анализ зависимости между температурой воздуха и продажами мороженого в киоске: при повышении температуры на 1 °C продажи увеличиваются в среднем на 15 порций. Уравнение регрессии: продажи = 50 + 15 × температура, где 50 – базовые продажи при 0 °C
MULTIPLE LINEAR REGRESSION	Регрессия с несколькими независимыми переменными (несколькими признаками). Множественная регрессия описывается уравнением $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$, позволяя учитывать влияние множества факторов одновременно	Прогнозирование стоимости доставки курьером учитывает несколько признаков: расстояние (км), вес заказа (кг), время суток (ч), день недели, погодные условия. Уравнение регрессии: стоимость = 100 + 15 × расстояние + 5 × вес + 20 × ночной_тариф + 10 × выходной
REGRESSION COEFFICIENTS	Параметры модели ($\beta_1, \beta_2, \dots, \beta_n$), определяющие вклад каждой независимой переменной в предсказание зависимой переменной. Коэффициент показывает, на сколько единиц изменится целевая	В модели предсказания зарплаты разработчика коэффициенты регрессии показывают следующее: опыт работы (лет) имеет коэффициент +15 000 (каждый год опыта добавляет 15 000 руб.), знание английского дает +25 000,

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	переменная в случае изменения соответствующего предиктора на одну единицу при фиксированных остальных переменных	работа в Москве +30 000, удаленная работа –10 000. Базовая зарплата (<i>intercept</i>) составляет 40 000 руб.
INTERCEPT	Значение зависимой переменной при нулевых значениях всех независимых переменных; точка пересечения регрессионной прямой с осью Y . Свободный член представляет базовое значение, когда все предикторы равны нулю	В модели стоимости такси <i>intercept</i> , равный 150 руб., означает базовую стоимость подачи автомобиля, которая взимается независимо от расстояния поездки. Даже если расстояние равно 0 км (пассажир не поехал), минимальная плата составляет 150 руб.
RESIDUAL	Разница между наблюдаемым (фактическим) значением зависимой переменной и значением, предсказанным моделью: $e_i = y_i - \hat{y}_i$. Остатки показывают, насколько хорошо модель описывает каждое конкретное наблюдение; анализ остатков помогает проверить предположения модели	Модель предсказала время доставки 35 мин, но фактическая доставка заняла 42 мин. Остаток: $42 - 35 = +7$ мин (модель недооценила время). Для другой доставки остаток составил –3 мин (модель переоценила время). Средний остаток по всем доставкам должен быть близок к нулю
SUM OF SQUARED ERRORS, SSE	Сумма квадратов разностей между наблюдаемыми и предсказанными значениями: $SSE = \sum (y_i - \hat{y}_i)^2$ SSE является мерой общей ошибки модели; метод наименьших квадратов (OLS) находит коэффициенты регрессии, минимизирующие эту сумму	Сравнение 2 моделей предсказания цен на жилье: 1-я модель имеет $SSE = 2\,500\,000$, 2-я $SSE = 1\,800\,000$. 2-я модель лучше, так как сумма квадратов ошибок меньше – предсказания 2-й модели в среднем ближе к фактическим ценам

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
R-SQUARED, R^2	Доля дисперсии зависимой переменной, объясненная моделью; показатель качества подгонки модели. R^2 принимает значения от 0 до 1: $R^2 = 0,85$ означает, что 85 % вариации целевой переменной объясняются признаками модели, а 15 % остаются необъясненными (случайная ошибка)	Модель предсказания рейтинга фильмов на Кинопоиске имеет $R^2 = 0,72$: 72 % различий в оценках пользователей объясняются признаками (жанр, бюджет, рейтинг режиссера, актеры), а 28 % обусловлены субъективными факторами, которые модель не учитывает
ADJUSTED R-SQUARED	Модификация R^2 , скорректированная на число предикторов в модели. В отличие от обычного R^2 , который всегда растет при добавлении новых признаков, Adjusted R^2 штрафует за добавление бесполезных предикторов, не улучшающих модель существенно	При добавлении в модель предсказания зарплат признака «любимый цвет программиста» обычный R^2 вырос с 0,78 до 0,781 (незначительно), но Adjusted R^2 снизился с 0,76 до 0,758 – это сигнал, что новый признак не несет полезной информации и только усложняет модель
MEAN SQUARED ERROR, MSE	Среднее значение квадратов ошибок: $MSE = \left(\frac{1}{n}\right) \sum (y_i - y)^2.$ MSE широко используется как функция потерь (<i>loss function</i>) в регрессии; квадратичная природа усиливает штраф за большие ошибки	Сравнение 2 алгоритмов прогнозирования спроса: 1-й имеет MSE = 144, 2-й MSE = 100. 2-й алгоритм точнее: среднеквадратичная ошибка меньше. MSE измеряется в квадратах единиц целевой переменной (если спрос в штуках, MSE – в штуках ²)
ROOT MEAN SQUARED ERROR, RMSE	Квадратный корень из MSE: $RMSE = \sqrt{MSE}$. RMSE измеряется в тех же единицах, что и целевая переменная, что упрощает интерпретацию. RMSE показывает «типичную»	Модель предсказания времени доставки имеет $RMSE = 8$ мин: это означает, что типичное отклонение предсказания от фактического времени составляет около 8 мин. Если среднее время

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	величину ошибки предсказания	доставки – 35 мин, то большинство предсказаний попадают в диапазон ± 8 мин от истины
MEAN ABSOLUTE ERROR, MAE	Среднее значение абсолютных ошибок: $MAE = \frac{1}{n} \sum y_i - \hat{y}_i ,$ где y_i – фактическое (истинное) значение целевой переменной; $\hat{y}_i$ – предсказанное моделью значение; n – количество наблюдений. MAE менее чувствительна к выбросам, чем MSE, так как не возводит ошибки в квадрат; MAE также измеряется в исходных единицах целевой переменной	2 модели предсказания цен на квартиры: 1-я имеет MAE = 350 000 руб., MSE = 200 млрд руб.; 2-я MAE = 380 000 руб., MSE = 180 млрд руб. 1-я модель в среднем ошибается меньше (MAE), но 2-я реже делает катастрофически большие ошибки (MSE)
OVERFITTING	Ситуация, когда модель слишком хорошо подстраивается под обучающие данные, включая их шум и случайные флуктуации, и теряет способность обобщаться на новые данные. Переобученная модель показывает отличные результаты на тренировочных данных, но плохие на тестовых	Модель распознавания кошек и собак выучила всю 1000 тренировочных фотографий наизусть (точность 100 %), но на новых фотографиях точность упала до 60 %. Модель запомнила конкретные изображения, а не научилась распознавать общие признаки животных – это переобучение
UNDERFITTING	Ситуация, когда модель слишком проста для данных и не может уловить основные закономерности. Недообученная модель показывает плохие результаты как на обучающих, так и на тестовых	Попытка описать сложные сезонные продажи мороженого (пики летом, спады зимой, рост по выходным) прямой линией дает точность 45 % – модель слишком проста и не учитывает сезонность, день недели, погоду.

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	данных; модель имеет высокое смещение (<i>bias</i>)	Нужно использовать более сложную модель с полиномиальными признаками
BIAS-VARIANCE TRADEOFF	Фундаментальный компромисс в машинном обучении: уменьшение смещения (<i>bias</i>) обычно увеличивает дисперсию (<i>variance</i>) и наоборот. Идеальная модель балансирует между недообучением (высокое смещение) и переобучением (высокая дисперсия)	Простая линейная регрессия для предсказания цен на жилье имеет высокое смещение (не улавливает нелинейности), но низкую дисперсию (стабильна на разных выборках). Сложная нейросеть имеет низкое смещение (точно описывает тренировочные данные), но высокую дисперсию (результаты сильно зависят от конкретной выборки). Оптимальна модель средней сложности (например, градиентный бустинг)
POLYNOMIAL REGRESSION	Регрессия с полиномиальными признаками (степенями исходных переменных), позволяющая моделировать нелинейные зависимости. Полиномиальная регрессия остается линейной по параметрам, но включает признаки x^2 , x^3 и т.д.	Зависимость между скоростью автомобиля и расходом топлива нелинейна: на низких скоростях расход растет медленно, на средних – оптимально, на высоких – резко увеличивается. Полиномиальная регрессия 2-й степени (с признаками «скорость» и «скорость ² ») точно описывает эту U-образную зависимость
REGULARIZATION	Метод предотвращения переобучения путем добавления штрафного члена к функции потерь, ограничивающего величину коэффициентов модели. Регуляризация уменьшает сложность модели, улучшая способность к обобщению	При обучении модели предсказания оттока клиентов добавлена L2-регуляризация (<i>Ridge</i>). Коэффициенты модели стали меньше по модулю (от 0,01 до 0,5 вместо значений от -5 до +10 без регуляризации), точность на тестовой выборке выросла с 78 до 84 % – модель лучше обобщает

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
RIDGE REGRESSION	Метод регуляризации L2, добавляющий к функции потерь штраф за сумму квадратов коэффициентов: $\lambda \cdot \sum_{j=1}^p \beta_j^2$. Ridge-регрессия сжимает коэффициенты, но не обнуляет их полностью; все признаки остаются в модели	В задаче предсказания рейтинга фильма по 100 признакам (жанры, актеры, режиссер, бюджет и т.д.) Ridge-регрессия уменьшила все коэффициенты пропорционально, но сохранила все признаки. Точность на валидации 82 против 76 % без регуляризации – переобучение устранено
LASSO REGRESSION	Метод регуляризации L1, добавляющий штраф за сумму абсолютных значений коэффициентов: $\lambda \cdot \sum_{j=1}^p \beta_j$. Lasso может обнулять некоторые коэффициенты, выполняя автоматический отбор признаков (<i>feature selection</i>)	При предсказании стоимости страхования автомобиля из 50 признаков Lasso обнулil 23 коэффициента (оставил только значимые: возраст водителя, стаж, марку авто, мощность двигателя, регион, историю аварий). Модель стала более простой и интерпретируемой, точность не ухудшилась
ELASTIC NET	Комбинация L1- и L2-регуляризации, объединяющая преимущества Ridge и Lasso. Elastic Net особенно полезен при высокой корреляции признаков (мультиколлинеарности) и в случае, когда число признаков превышает число наблюдений	В геномном анализе для предсказания заболевания по 20 000 генов (признаков) при всего 500 пациентах Elastic Net отобрала 150 значимых генов (L1-компонент) и стабилизировала коэффициенты (L2-компонент). Точность классификации – 89 против 76 % у чистого Lasso
MULTI-COLLINEARITY	Высокая корреляция между независимыми переменными (признаками) в модели регрессии. Мультиколлинеарность затрудняет интерпретацию отдельных коэффициентов (невозможно отделить	В модели предсказания цены квартиры признаки «площадь в квадратных метрах» и «площадь в квадратных футах» имеют корреляцию 0,99 (полная мультиколлинеарность). Коэффициенты становятся

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	влияние коррелирующих признаков) и увеличивает дисперсию оценок	неустойчивыми: при добавлении/удалении других признаков коэффициенты меняют знак. Нужно удалить один из признаков
HOMO-SCEDASTICITY	Постоянство дисперсии ошибок (остатков) модели across всех уровней независимых переменных; одно из ключевых предположений линейной регрессии. При гомоскедастичности разброс остатков одинаков для всех предсказанных значений	На графике остатков модели предсказания зарплат видно следующее: разброс точек (остатков) примерно одинаков для всех уровней предсказанных зарплат (от 40 до 300 тыс. руб.) – дисперсия ошибок постоянна, предположение гомоскедастичности выполняется, стандартные ошибки коэффициентов надежны
HETERO-SCEDASTICITY	Непостоянство дисперсии ошибок: разброс остатков меняется в зависимости от значений предикторов или предсказанных значений. Гетероскедастичность приводит к неэффективным оценкам и смещенным стандартным ошибкам, что искажает проверку гипотез	Модель предсказания расходов домохозяйств показывает следующее: для бедных семей (доход 30 тыс. руб.) остатки маленькие (± 2 тыс. руб.), для богатых (доход 300 тыс.) разброс огромный (± 50 тыс. руб.). Это гетероскедастичность: дисперсия ошибок растет с доходом. Нужна робастная оценка стандартных ошибок или трансформация данных

3.2. Probabilistic Classification and Evaluation of Models

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
CLASSIFICATION	Задача предсказания категориальной метки класса для нового наблюдения на основе обучающих данных. В отличие от регрессии (прогноз числа), классификация относит объект к одному из заранее определенных классов (категорий)	Система безопасности банка классифицирует каждую транзакцию: мошенничество или легитимная. Почтовый сервис классифицирует письма: спам или не спам. Медицинская система диагностирует: здоров/болен/требуется дополнительные анализы
BINARY CLASSIFICATION	Классификация с 2 классами (положительный и отрицательный). Бинарная классификация – наиболее распространенный тип задач: спам/не спам, болен/здоров, дефолт/платежеспособен, клик/нет клика	Модель определяет, кликнет ли пользователь на рекламный баннер (класс 1) или игнорирует (класс 0). На основе признаков (история кликов, время на сайте, устройство, источник трафика) алгоритм предсказывает вероятность клика и относит пользователя к одному из 2 классов
MULTICLASS CLASSIFICATION	Классификация с 3 и более классами. Многоклассовая классификация требует специальных алгоритмов или стратегий «один против всех» (one-vs-rest) или «один против одного» (one-vs-one)	Распознавание рукописных цифр MNIST: изображение цифры (28×28 пикселей) классифицируется как один из 10 классов (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). Классификация новостей по темам: спорт, политика, экономика, культура, технологии (5 классов)
LOGISTIC REGRESSION	Статистическая модель для бинарной классификации, предсказывающая вероятность принадлежности к положительному классу.	Модель предсказывает вероятность одобрения кредитной заявки: на основе признаков (доход, кредитная история,

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	Логистическая регрессия использует сигмоидную функцию для преобразования линейной комбинации признаков в вероятность от 0 до 1	возраст, сумма кредита) логистическая регрессия выдает вероятность 0,73 (73 %). Если порог 0,5, заявка одобряется; если порог 0,8 – отклоняется
SIGMOID FUNCTION	S-образная функция, отображающая любые входные значения в диапазоне от 0 до 1: $\sigma(z) = 1/(1 + e^{-z})$. Сигмоида преобразует линейную комбинацию признаков в вероятность, что делает ее идеальной для логистической регрессии и нейронных сетей	Логистическая регрессия вычислила, что $z = 2,5$ (линейная комбинация признаков). Сигмоида $\sigma(2,5) = 0,92$ – модель предсказывает 92%-ю вероятность положительного класса. Если $z = -1,5$, $\sigma(-1,5) = 0,18$, вероятность всего 18 %, объект скорее относится к отрицательному классу
DECISION BOUNDARY	Граница в пространстве признаков, разделяющая классы; поверхность, по одну сторону от которой модель предсказывает один класс, по другую – другой класс. В логистической регрессии граница линейна (прямая линия или гиперплоскость)	В задаче классификации ирисов по 2 признакам (длина и ширина лепестка) логистическая регрессия построила линейную границу: если $2 \times \text{длина} + 3 \times \text{ширина} > 15$, предсказываем вид Iris Virginica, иначе Iris Setosa. Точки по разные стороны прямой относятся к разным классам
PROBABILITY THRESHOLD	Пороговое значение вероятности для отнесения объекта к положительному классу. По умолчанию порог обычно устанавливается на уровне 0,5, но может быть смещен в зависимости от стоимости ошибок разных типов	Модель диагностики рака молочной железы выдает вероятность 0,47. При стандартном пороге 0,5 результат «здоров», но так как ложноотрицательный результат (пропуск рака) опасен, порог снижен до 0,3. Теперь

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
		вероятность 0,47 классифицируется как «требуется дополнительное обследование»
CONFUSION MATRIX	Таблица размером $K \times K$ (где K – число классов) для оценки качества классификации, показывающая количество истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных предсказаний	Модель диагностики пневмонии по рентгену на 100 пациентах показала следующее: 45 истинно больных (<i>True Positive</i>), 40 истинно здоровых (<i>True Negative</i>), 5 ложноположительных (здоровых назвали больными – <i>False Positive</i>), 10 ложноотрицательных (больных назвали здоровыми – <i>False Negative</i>). Матрица ошибок помогает рассчитать точность, полноту, специфичность
TRUE POSITIVE, TP	Правильно предсказанный положительный класс: модель предсказала положительный класс, и это соответствует действительности	Система обнаружения мошенничества правильно определила 247 мошеннических транзакций из 300 реальных (TP = 247). Это хорошие результаты: большинство реальных случаев мошенничества обнаружено
TRUE NEGATIVE, TN	Правильно предсказанный отрицательный класс: модель предсказала отрицательный класс, и это соответствует действительности	Спам-фильтр правильно пропустил 1 850 легитимных писем в папку «Входящие» (TN = 1 850). Важные письма не потеряны, пользователи довольны работой фильтра
FALSE POSITIVE, FP	Ошибочно предсказанный положительный класс (ошибка I рода): модель предсказала положитель-	Тест на беременность показал положительный результат, но женщина не беременна (FP).

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	ный класс, но на самом деле объект относится к отрицательному классу	В спам-фильтрации FP – это когда важное рабочее письмо было ошибочно помечено как спам и попало в папку «Спам» (пользователь может его не увидеть)
FALSE NEGATIVE, FN	Ошибочно предсказанный отрицательный класс (ошибка II рода): модель предсказала отрицательный класс, но на самом деле объект относится к положительному классу	Тест на COVID-19 показал отрицательный результат у зараженного человека (FN = ложноотрицательный). Человек считает себя здоровым, не изолируется и заражает других. В медицине FN часто опаснее FP
ACCURACY	Доля правильных предсказаний среди всех предсказаний: $Accuracy = (TP + TN) / (TP + TN + FP + FN)$. Accuracy проста для интерпретации, но может вводить в заблуждение при несбалансированных классах	Модель диагностики редкого заболевания (1 % больных) просто предсказывает всем «здоров» и получает Accuracy = 99 %. Но это бесполезная модель: она не находит ни одного больного! Для несбалансированных данных нужны Precision, Recall, F1-Score
PRECISION	Доля истинно положительных среди всех предсказанных положительных: $Precision = TP / (TP + FP)$. Precision отвечает на вопрос «Среди объектов, которые модель назвала положительными, какая доля действительно положительна?»	Спам-фильтр пометил 120 писем как спам, из которых 108 действительно спам (TP = 108), а 12 – важные письма (FP = 12). $Precision = 108 / 120 = 0,90$ (90 %). Высокая Precision означает, что, когда фильтр говорит: «Спам», он обычно прав
RECALL	Доля истинно положительных среди всех фактически положительных:	В базе 300 мошеннических транзакций, модель обнаружила 247 (TP = 247), пропустила

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	Recall = $TP / (TP + FN)$. Recall отвечает на вопрос: «Какую долю реальных положительных объектов нашла модель?»	53 (FN = 53). Recall = $= 247 / 300 = 0,82$ (82 %). Модель нашла 82 % реального мошенничества. Высокий Recall критичен, когда пропуск положительного случая опасен (диагностика рака)
SENSITIVITY	Синоним Recall (полноты); доля истинно положительных среди всех фактически положительных. Sensitivity измеряет способность модели обнаруживать положительные случаи	Тест на ВИЧ имеет Sensitivity = 0,995 (99,5 %): из 1000 зараженных тест правильно обнаружит 995 человек, и только 5 получают ложноотрицательный результат. Высокая чувствительность критична для скрининговых тестов
SPECIFICITY	Доля истинно отрицательных среди всех фактически отрицательных: $\text{Specificity} = \frac{TN}{TN + FP}.$ Specificity измеряет способность модели правильно идентифицировать отрицательные случаи	Тест на аллергию имеет Specificity = 0,92 (92 %): из 100 здоровых людей тест правильно определит 92 как не имеющих аллергии, а 8 получают ложноположительный результат. Высокая специфичность важна, когда ложная тревога ведет к дорогостоящим дополнительным обследованиям
F1-SCORE	Гармоническое среднее точности (Precision) и полноты (Recall): $F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}.$ F1-Score полезен, когда нужен баланс между Precision и Recall, особенно при несбалансированных классах	Модель A: Precision = 0,95; Recall = 0,60; F1 = 0,74. Модель B: Precision = 0,80; Recall = 0,85; F1 = 0,82. Хотя у модели A выше Precision, модель B лучше по F1 – она обеспечивает лучший баланс между

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
		нахождением положительных случаев и минимизацией ложных срабатываний
ROC CURVE	График зависимости True Positive Rate (Recall) от False Positive Rate при различных пороговых значениях. ROC-кривая визуализирует компромисс между обнаружением положительных случаев и ложными тревогами	При пороге 0,9: TPR = 0,50; FPR = 0,05. При пороге 0,7: TPR = 0,75; FPR = 0,15. При пороге 0,5: TPR = 0,90; FPR = 0,30. При пороге 0,3: TPR = 0,98; FPR = 0,60. ROC-кривая соединяет эти точки; чем ближе кривая к верхнему левому углу, тем лучше модель
AUC (AREA UNDER ROC CURVE)	Площадь под ROC-кривой; интегральная метрика качества классификатора. AUC принимает значения от 0,5 до 1,0: AUC = 0,5 означает случайное угадывание, AUC = 1,0 – идеальную классификацию. AUC интерпретируется как вероятность того, что модель оценит случайно выбранный положительный пример выше, чем случайно выбранный отрицательный	Модель A имеет AUC = 0,92 – отличное качество: в 92 % случаев модель правильно ранжирует положительный пример выше отрицательного. Модель B: AUC = 0,68 – слабая модель, лишь немного лучше случайного угадывания. Модель C: AUC = 0,50 – бесполезная модель (случайный классификатор)
TRUE POSITIVE RATE, TPR	Доля правильно предсказанных положительных случаев среди всех фактически положительных: $TPR = TP / (TP + FN)$. TPR совпадает с Recall и Sensitivity	Из 200 пациентов с диагнозом «пневмония» модель правильно определила 176 (TP = 176), пропустила 24 (FN = 24). $TPR = 176 / 200 = 0,88$ (88 %). Модель обнаружила 88 % реальных случаев пневмонии
FALSE POSITIVE RATE, FPR	Доля ошибочно предсказанных положительных случаев среди всех фактически отрицательных:	Из 1000 здоровых людей тест дал 80 ложноположительных результатов (FP = 80), 920 –

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	$FPR = FP/(FP + TN)$. FPR показывает, как часто модель «ложно тревожится»	правильных отрицательных ($TN = 920$). $FPR = 80/1000 = 0,08$ (8 %). Тест ошибочно диагностирует заболевание у 8 % здоровых людей
CROSS-VALIDATION	Метод оценки обобщающей способности модели путем многократного разбиения данных на обучающую и валидационную выборки. Кросс-валидация дает более надежную оценку качества, чем однократное разбиение на train/test	Вместо одного разбиения 80/20 данные делятся на 5 частей. Модель обучается 5 раз: каждый раз одна из частей используется как тестовая, остальные 4 – как обучающие. Получаем 5 оценок точности: 0,82; 0,79; 0,84; 0,81; 0,83. Средняя точность $0,82 \pm 0,02$ – надежная оценка качества
K-FOLD CROSS-VALIDATION	Кросс-валидация с разделением данных на k равных частей. Наиболее распространена 10-кратная кросс-валидация ($k = 10$), обеспечивающая хороший баланс между вычислительной стоимостью и надежностью оценки	Набор данных из 5000 записей делится на 10 фолдов по 500 записей. Модель обучается 10 раз: первый раз фолд 1 – тест, фолды 2–10 – тренировка; второй раз фолд 2 – тест, остальные – тренировка, и т.д. Итоговая метрика – среднее по 10 запускам
LEAVE-ONE-OUT, LOO	Кросс-валидация, где каждый объект поочередно становится тестовым, а все остальные – обучающими (частный случай k -fold, где $k = n$, числу объектов). LOO использует максимум данных для обучения, но вычислительно дорога	Набор из 100 пациентов: модель обучается 100 раз. Первый раз: пациент 1 – тест, пациенты 2–100 – тренировка. Второй раз: пациент 2 – тест, пациенты 1, 3–100 – тренировка. И так 100 раз. LOO дает наименее смещенную оценку, но требует 100 обучений модели

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
TRAINING SET	Набор данных для обучения (подгонки) модели; данные, на которых алгоритм «учится», подбирая параметры (веса, коэффициенты). Обычно составляет 60–80 % всех данных	Из 10 000 размеченных изображений кошек и собак 8 000 (80 %) выделены в тренировочный набор. На этих 8 000 изображений нейросеть обучается 50 эпох, корректируя веса для минимизации функции потерь
VALIDATION SET	Набор данных для настройки гиперпараметров и выбора лучшей конфигурации модели; данные, используемые для промежуточной оценки в процессе обучения. Валидационная выборка помогает предотвратить переобучение	Из 10 000 изображений 8 000 – тренировка, 1 000 – валидация, 1 000 – тест. В процессе обучения после каждой эпохи проверяем точность на валидации. Если точность на тренировке растет, а на валидации падает, началось переобучение, нужно остановить обучение (<i>early stopping</i>)
TEST SET	Набор данных для финальной оценки модели; данные, которые не используются ни для обучения, ни для настройки гиперпараметров. Тестовая выборка должна использоваться только один раз для оценки финальной производительности модели	После завершения обучения и настройки гиперпараметров на тренировке и валидации, финальная модель оценивается на тестовой выборке (1 000 изображений, которые модель «не видела»). Точность на тесте 87 % – это реалистичная оценка того, как модель будет работать на новых данных
HYPER-PARAMETERS	Параметры, настраиваемые до обучения модели (в отличие от параметров модели, которые обучаются на данных). Примеры: скорость обучения (<i>learning rate</i>), число деревьев в случайном лесе,	Градиентный бустинг имеет гиперпараметры: $\text{learning}_{\text{rate}} = 0,1$; $n_{\text{estimators}} = 100$; $\text{max}_{\text{depth}} = 5$; $\text{min}_{\text{samples}_{\text{split}}} = 10$. Эти значения задаются

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	степень полинома, коэффициент регуляризации	до обучения. Подбор гиперпараметров нашел лучшие значения: $\text{learning}_{\text{rate}} = 0,05$; $n_{\text{estimators}} = 200$; $\text{max}_{\text{depth}} = 7$ – точность выросла с 82 до 87 %
GRID SEARCH	Метод подбора гиперпараметров полным перебором: систематическая проверка всех комбинаций заданных значений гиперпараметров с использованием кросс-валидации для оценки каждой комбинации	Для SVM задаем сетку: $C = [0,1; 1; 10]$, $\text{gamma} = [0,001; 0,01; 0,1]$, $\text{kernel} = [\text{linear}, \text{rbf}]$. Всего $3 \cdot 3 \cdot 2 = 18$ комбинаций. Каждая комбинация оценивается 5-кратной кросс-валидацией (90 обучений модели). Лучшая комбинация: $C = 10$, $\text{gamma} = 0,01$; $\text{kernel} = \text{rbf}$ с точностью 91 %
BAYESIAN CLASSIFICATION	Классификация на основе теоремы Байеса, обновляющая априорные убеждения с учетом наблюдаемых данных (<i>evidence</i>). Байесовские методы дают вероятностную интерпретацию предсказаний	Априорная вероятность заболевания 1 % ($P(\text{болезнь}) = 0,01$). Тест дал положительный результат. По теореме Байеса вычисляем апостериорную вероятность: $P(\text{болезнь} \text{положительный_тест}) = 0,32$ (32 %). Несмотря на положительный тест, вероятность болезни всего 32 % из-за ложноположительных результатов теста
NAIVE BAYES	Классификатор с предположением о независимости признаков при условии класса (условная независимость). Несмотря на «наивное» предположение, Naive Bayes удивительно эффективен для	Спам-фильтр анализирует email: признаки – наличие слов «выигрыш» (вероятность в спаме 0,4, в не спаме 0,01), «бесплатно» (0,3 vs 0,02), «срочно» (0,25 vs 0,03). Naive Bayes перемножает

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	текстовой классификации и спам-фильтрации	вероятности (предполагая независимость слов) и получает следующее: $P(\text{спам} \text{слова}) = 0,89$; $P(\text{не спам} \text{слова}) = 0,11$. Письмо классифицируется как спам
PRIOR PROBABILITY	Вероятность гипотезы (класса) до учета данных; начальная степень уверенности в гипотезе, основанная на предыдущих знаниях или частоте классов в данных	В задаче диагностики редкого заболевания априорная вероятность $P(\text{болен}) = 0,01$ (1 % населения болеет), $P(\text{здоров}) = 0,99$. Даже при положительном тесте (который имеет точность 95 %) апостериорная вероятность болезни остается низкой из-за малой априорной вероятности
POSTERIOR PROBABILITY	Вероятность гипотезы (класса) после учета данных; обновленная вероятность, вычисленная по теореме Байеса на основе априорной вероятности и правдоподобия данных	Априорная вероятность спама $P(\text{спам}) = 0,30$ (30 % писем – спам). Письмо содержит слово «выигрыш», которое встречается в 40 % спама и 1 % не спама. По Байесу $P(\text{спам} \text{«выигрыш»}) = 0,95$. После учета наличия слова апостериорная вероятность спама выросла с 30 до 95 %
LIKELIHOOD	Вероятность данных при заданной гипотезе; вероятность наблюдать имеющиеся данные, если гипотеза верна. В теореме Байеса: $P(\text{данные} \text{гипотеза})$	Гипотеза H : монета честная ($P(\text{орел}) = 0,5$). Данные D : 8 орлов из 10 бросков. Правдоподобие $P(D H) = C(10,8) \cdot 0,5^8 \cdot 0,5^2 = 0,044$ (4,4 %). Низкое правдоподобие указывает на то, что данные маловероятны при гипотезе о честной монете – возможно, монета нечестная

SECTION 4

ADVANCED TOPICS: UNSUPERVISED TECHNIQUES, SCALABILITY, AND AI INTEGRATION

4.1. Statistical Approaches to Unsupervised Learning

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
UNSUPERVISED LEARNING	Тип машинного обучения, при котором алгоритм работает с неразмеченными данными (без известных правильных ответов). Цель – обнаружить скрытые паттерны, структуры или взаимосвязи в данных самостоятельно, без внешнего руководства	Анализ поведения пользователей мобильного приложения без заранее известных категорий: алгоритм самостоятельно обнаруживает, что есть группы пользователей «утренние активные», «вечерние пассивные», «выходного дня», «постоянные», хотя мы заранее не знали о существовании таких сегментов
CLUSTERING	Группировка объектов по сходству их характеристик; задача разбиения набора данных на группы (кластеры) таким образом, чтобы объекты внутри одного кластера были похожи друг на друга, а объекты из разных кластеров отличались	Сегментация клиентов Интернет-магазина по покупательскому поведению: кластер 1 – «частые покупатели с низким средним чеком»; кластер 2 – «редкие покупатели с высоким чеком»; кластер 3 – «сезонные покупатели»; кластер 4 – «уходящие клиенты». Для каждого кластера разрабатывается своя маркетинговая стратегия

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
K-MEANS CLUSTERING	Итеративный метод кластеризации, который делит данные на K кластеров путем минимизации внутрикластерной дисперсии (суммы квадратов расстояний от точек до центроидов их кластеров). Алгоритм: 1) случайно инициализировать K центроидов; 2) назначить точки ближайшим центроидам; 3) пересчитать центроиды; 4) повторять шаги 2, 3 до сходимости	Кластеризация 10 000 статей новостного сайта на 5 тематических групп: алгоритм K -means анализирует векторные представления текстов (TF-IDF или word2vec) и автоматически группирует статьи в кластеры: политика, спорт, экономика, культура, технологии – без заранее известных меток категорий
HIERARCHICAL CLUSTERING	Метод кластеризации, строящий иерархическую древовидную структуру (дендрограмму) вложенных кластеров. Бывает агломеративная (снизу вверх: каждый объект – отдельный кластер, затем кластеры объединяются) и дивизивная (сверху вниз: все объекты в одном кластере, затем разделяются)	Иерархическая кластеризация 100 видов животных по генетическим признакам: дендрограмма показывает эволюционное родство – на верхнем уровне разделяются млекопитающие, птицы, рептилии; внутри млекопитающих – хищники, травоядные, всеядные; внутри хищников – кошачьи, псовые и т.д.
DENDROGRAM	Древовидная диаграмма, визуализирующая результаты иерархической кластеризации. Показывает последовательность объединения или разделения кластеров и расстояния (сходство) между ними. Высота ветвей дендрограммы отражает степень различия между кластерами	Дендрограмма кластеризации 50 компаний по финансовым показателям демонстрирует следующее: на высоте 10 единиц все компании в одном кластере; на высоте 5 – разделение на 3 крупных кластера (технологические, промышленные, финансовые); на высоте 2 – внутри технологических выделяются софтверные и аппаратные компании

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE)	Алгоритм кластеризации на основе плотности, который находит кластеры произвольной формы и выявляет выбросы (шум). DBSCAN группирует вместе точки, находящиеся близко друг к другу (в пределах радиуса ϵ), и помечает изолированные точки как шум. Параметры: ϵ (радиус окрестности) и minPts (минимальное число точек)	Обнаружение аномалий в геолокации: DBSCAN анализирует координаты транзакций клиента. Основная масса точек образует плотный кластер (дом – работа – магазины), но несколько транзакций в другом городе/стране оказываются изолированными точками (шум) – это аномалии, возможное мошенничество
CENTROID	Центр кластера; точка, представляющая среднее положение всех точек, принадлежащих кластеру. В K -means центроид вычисляется как среднее арифметическое координат всех точек кластера по каждому признаку	После кластеризации клиентов на 4 группы центроиды кластеров показывают типичные характеристики: центроид кластера 1 – возраст 25 лет, доход 40 тыс., покупки 2 раза в месяц; центроид кластера 2 – возраст 45 лет, доход 120 тыс., покупки 8 раз в месяц
SILHOUETTE SCORE	Мера качества кластеризации, показывающая, насколько хорошо объекты разделены на кластеры. Для каждой точки вычисляется: a – среднее расстояние до точек своего кластера; b – среднее расстояние до точек ближайшего соседнего кластера: $\text{Silhouette} = \frac{b - a}{\max(a, b)}$ Значения от -1 до $+1$: чем выше, тем лучше кластеризация	Сравнение кластеризации с $K = 3$ и $K = 5$: при $K = 3$ средний Silhouette Score = 0,65 (хорошее разделение), при $K = 5$ Silhouette Score = 0,42 (кластеры перекрываются). Вывод: 3 кластера оптимальнее, чем 5

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
ELBOW METHOD	Эвристический метод определения оптимального числа кластеров K в K -means. Строится график зависимости внутрикластерной дисперсии (<i>inertia</i>) от числа кластеров. Оптимальное K находится в точке «излома» (<i>elbow</i>), где добавление новых кластеров перестает существенно уменьшать дисперсию	График <i>inertia</i> от $K = 1$ до $K = 10$ показывает резкое снижение до $K = 4$ (<i>inertia</i> падает с 1000 до 300), затем плавное снижение ($K = 5$: 280, $K = 6$: 265). Точка излома на $K = 4$ – оптимальное число кластеров для данных
DIMENSIONALITY REDUCTION	Процесс уменьшения числа признаков (переменных) в данных при сохранении важной информации и структуры. Снижение размерности помогает визуализировать многомерные данные, уменьшить вычислительные затраты и избежать проклятия размерности	Исходные данные содержат 100 признаков о клиентах (демография, поведение, предпочтения). После PCA (Principal Component Analysis) данные проецируются на 3 главные компоненты, сохраняющие 85 % дисперсии. Теперь данные можно визуализировать в 3D и быстрее обучать модели
PRINCIPAL COMPONENT ANALYSIS, PCA	Метод снижения размерности, который преобразует исходные коррелированные признаки в новый набор некоррелированных признаков (главных компонент), упорядоченных по количеству объясняемой дисперсии. Первая компонента объясняет наибольшую дисперсию, вторая – следующую по величине и т.д.	Анализ антропометрических данных (рост, вес, объем груди, талии, бедер, длина рук и ног – 15 признаков). PCA показывает, что первая главная компонента (общий размер тела) объясняет 60 % дисперсии, вторая (пропорции) – 20 %, третья – 10 %. Достаточно первых 3 компонент вместо 15 исходных признаков
T-SNE	Метод визуализации многомерных данных, проецирующий данные в 2D-или 3D-пространство	Визуализация 10 000 изображений рукописных цифр MNIST (каждое

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	с сохранением локальной структуры (близкие точки в исходном пространстве остаются близкими). <i>t</i> -SNE особенно хорош для визуализации кластеров	изображение – 784 признака). <i>t</i> -SNE проецирует данные в 2D: на графике четко видны 10 отдельных кластеров, соответствующих цифрам 0–9, хотя исходные данные были 784-мерными
UMAP (UNIFORM MANIFOLD APPROXIMATION AND PROJECTION)	Метод снижения размерности, сохраняющий как локальную, так и глобальную структуру данных. UMAP быстрее <i>t</i> -SNE и лучше масштабируется на большие наборы данных, при этом часто сохраняя больше глобальной структуры	Анализ генетических данных 50 000 человек (20 000 генов). UMAP проецирует данные в 2D за 10 мин (<i>t</i> -SNE потребовалось бы несколько часов). На карте видны кластеры по этническому происхождению (европейцы, азиаты, африканцы) и градиенты внутри групп
EIGENVALUE	Характеристическое число матрицы ковариации в PCA, показывающее количество дисперсии данных, захваченное соответствующей главной компонентой. Чем больше собственное значение, тем больше информации (вариативности) сохраняет компонента	Матрица ковариации 15 признаков имеет собственные значения: $\lambda_1 = 8,5$; $\lambda_2 = 3,2$; $\lambda_3 = 1,8$; $\lambda_4 = 0,9...$ Первая главная компонента ($\lambda_1 = 8,5$) объясняет $8,5/15 = 57\%$ общей дисперсии, первые 3 компоненты вместе: $(8,5 + 3,2 + 1,8)/15 = 90\%$
EIGENVECTOR	Вектор, определяющий направление главной компоненты в пространстве признаков. Показывает, как исходные признаки комбинируются для формирования новой компоненты. Собственный вектор нормализован (единичной длины)	Первый собственный вектор PCA антропометрических данных: [0,35; 0,36; 0,34; 0,35; 0,33; ...] – все коэффициенты положительные и примерно равные, значит, первая компонента – это «общий размер тела»

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
		(усреднение всех измерений). Второй вектор: [0,1; -0,5; 0,2; -0,6; 0,3; ...] – чередование знаков показывает «пропорции» (соотношение разных частей тела)
EXPLAINED VARIANCE	Доля общей дисперсии данных, сохраняемая компонентой или набором компонент. Показывает процент дисперсии, объясненный каждой главной компонентой	РСА 20 признаков: первая компонента объясняет 45 % дисперсии, вторая – 20 %, третья – 12 %, четвертая – 8 %. Первые 4 компоненты вместе объясняют 85 % общей вариативности данных – можно отбросить остальные 16 компонент с потерей всего 15 % информации
ASSOCIATION RULES	Правила, описывающие взаимосвязи между переменными (элементами) в виде «если X , то Y ». Ассоциативные правила используются для анализа рыночной корзины и обнаружения часто совместно встречающихся элементов	Анализ чеков супермаркета выявляет следующее правило: {пиво, чипсы} $\rightarrow$ $\rightarrow$ {арахис} с поддержкой 5 %, достоверностью 70 % и lift 3,2. Это означает, что 5 % всех покупок содержат все 3 товара; из покупателей, взявших пиво и чипсы, 70 % также берут арахис; арахис покупают с пивом и чипсами в 3,2 раза чаще, чем случайно
SUPPORT	Частота появления набора элементов в данных; доля транзакций (наблюдений), содержащих данный набор элементов.	В базе 10 000 покупок: 500 содержат {молоко, хлеб}, support = $= 500/10\,000 = 0,05$ (5 %). Правило {молоко} $\rightarrow$ {хлеб} имеет

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	$\text{Support} = P(X \cup Y)$ – вероятность совместного появления элементов	поддержку 5 % – это относительно частая комбинация товаров
CONFIDENCE	Условная вероятность появления элемента Y при наличии элемента X . Показывает, насколько часто правило оказывается верным: $\text{Confidence} = P(Y X) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$	Из 1000 покупок: 200 содержат молоко, 150 содержат {молоко, хлеб}. $\text{Confidence}(\{\text{молоко}\} \rightarrow \{\text{хлеб}\}) = 150/200 = 0,75$ (75 %). Это означает: 75 % покупателей, взявших молоко, также берут хлеб
LIFT	Мера силы ассоциативного правила, показывающая, насколько чаще элементы встречаются вместе по сравнению со случайным совпадением: $\text{Lift} = \text{Confidence}/P(Y) = P(Y X)/P(Y)$. $\text{Lift} > 1$ указывает на положительную связь; $\text{Lift} < 1$ – на отрицательную; $\text{Lift} = 1$ – на независимость	Правило {онлайн-курс по Python} $\rightarrow$ {онлайн-курс по Data Science} имеет $\text{Lift} = 4,5$. Это означает, что студенты, записавшиеся на курс по Python, покупают курс по Data Science в 4,5 раза чаще, чем случайные студенты. Сильная ассоциация, несмотря на низкую поддержку (всего 2 % студентов покупают оба курса)
ANOMALY DETECTION	Выявление нестандартных наблюдений (выбросов), которые значительно отличаются от большинства данных. Обнаружение аномалий используется для детекции мошенничества, сбоев оборудования, кибератак	Система мониторинга банковских транзакций: типичные операции клиента – покупки 500–5000 руб. в Москве. Внезапно появляется транзакция «снятие наличных 150 000 руб. в Бангкоке». Алгоритм обнаружения аномалий (<i>Isolation Forest</i>) помечает операцию как подозрительную (аномалию), банк блокирует карту и связывается с клиентом

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
GAUSSIAN MIXTURE MODELS, GMM	Вероятностная модель кластеризации, предполагающая, что данные сгенерированы смесью нескольких гауссовых (нормальных) распределений. В отличие от <i>K-means</i> (жесткая кластеризация), GMM дает мягкую кластеризацию – вероятность принадлежности к каждому кластеру	Кластеризация голосовых записей 3 дикторов: GMM моделирует каждый голос как многомерное нормальное распределение в пространстве акустических признаков. Для новой записи модель выдает: вероятность принадлежности диктору 1 = 0,75; диктору 2 = 0,15; диктору 3 = 0,10 – запись относится к первому диктору с уверенностью 75 %
EXPECTATION-MAXIMIZATION, EM	Итеративный алгоритм оценки параметров моделей со скрытыми переменными (например, GMM). EM чередует 2 шага: <i>E-step</i> (оценка вероятностей принадлежности к кластерам при текущих параметрах) и <i>M-step</i> (пересчет параметров для максимизации правдоподобия)	Обучение GMM с 4 компонентами: на <i>E-step</i> вычисляем для каждой точки вероятности принадлежности к 4 кластерам; на <i>M-step</i> пересчитываем средние, ковариации и веса компонент на основе этих вероятностей. Повторяем 50 итераций до сходимости – параметры стабилизируются

4.2. Statistics in Modern AI Systems and New Trends

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
DEEP LEARNING	Подраздел машинного обучения, использующий нейронные сети с множеством слоев (глубокие нейронные сети). Глубокое обуче-	Распознавание объектов на изображениях: сверточная нейронная сеть (CNN) с 50 слоями обучается на миллионах размеченных фото.

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	ние автоматически извлекает иерархические признаки из данных: от простых (края, текстуры) к сложным (объекты, концепции)	Первые слои учатся распознавать края и углы, средние – текстуры и простые формы, глубокие – части объектов (глаза, колеса), последние – целые объекты (лица, автомобили)
NEURAL NETWORK	Вычислительная система, вдохновленная биологическими нейронными сетями мозга; состоит из взаимосвязанных узлов (нейронов), организованных в слои. Каждый нейрон принимает входы, применяет взвешенную сумму и функцию активации, передает результат дальше	Классификация тональности отзывов: входной слой принимает вектор слов отзыва (1000 признаков), скрытый слой из 128 нейронов обрабатывает информацию, выходной слой из 2 нейронов выдает вероятности «позитивный»/«негативный». Сеть обучается на 10 000 размеченных отзывов
BACKPROPAGATION	Алгоритм обучения нейронных сетей, вычисляющий градиент функции потерь по всем весам сети с помощью цепного правила. Backpropagation распространяет ошибку от выходного слоя к входному, показывая, как изменить каждый вес для уменьшения ошибки	После прямого прохода через сеть (предсказание) вычисляется ошибка (<i>loss</i>). Backpropagation идет в обратном направлении: вычисляет градиент для весов выходного слоя, затем скрытого слоя, затем входного. Градиентный спуск обновляет веса: $w_new = w_old - learning_rate \times gradient$
GRADIENT DESCENT	Итеративный метод оптимизации, минимизирующий функцию потерь путем движения в направлении, противоположном градиенту (наискорейшего убывания). На каждой итерации параметры обнов-	Обучение линейной регрессии: начальная ошибка MSE = 100. Градиентный спуск с $learning_rate = 0,01$ делает 1000 шагов: после 100 итераций MSE = 45, после 500 итераций MSE = 12,

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	ляются: $\theta = \theta - \alpha \times \nabla J(\theta)$, где α – скорость обуче- ния; ∇J – градиент функции потерь	после 1000 итераций MSE = 8,5 (сходимость)
STOCHASTIC GRADIENT DESCENT, SGD	Вариант градиентного спуска, использующий случайные выборки (мини-батчи) вместо всего набора данных для вычисления гради- ента. SGD быстрее об- новляет параметры и может выходить из локальных минимумов благодаря шуму, но траектория менее ста- бильна	Обучение нейросети на 1 млн изображений: полный градиентный спуск требовал бы про- хода по всему 1 млн изображений для од- ного обновления (очень медленно). SGD ис- пользует мини-батчи по 32 изображения: после каждого 32-го обнов- ляет веса. За одну эпоху происходит 31 250 об- новлений вместо одного
BATCH SIZE	Число примеров, обра- батываемых за один шаг (перед обновле- нием весов). Влияет на требования к памяти, стабильность обучения и скорость сходимости. Малый батч – шумные обновления, но быстрая сходимость; большой батч – стабильные гра- диенты, но он медлен- нее и требует больше памяти	Эксперименты с размером батча для обучения CNN: batch_size = 32 – обуче- ние нестабильное (loss скачет), но быстро находит хорошее реше- ние; batch_size = 256 – стабильное плавное снижение потерь, опти- мальная скорость; batch_size = 2048 – очень стабильно, но требует много памяти и медленно сходится
EPOCH	Один полный проход по всему обучающему набору данных. Во время одной эпохи мо- дель видит каждый тре- нировочный пример ровно один раз (в SGD – в составе раз- личных батчей). Обуче- ние обычно требует множества эпох для сходимости	Обучение модели на 50 000 изображений с batch_size = 100: 1 эпоха = 500 итераций (50 000/100). После 1 эпохи точность на валидации – 45 %; после 10 эпох – 78 %; после 30 эпох – 85 %; после 50 эпох – 86 % (насыщение, дальнейшее обучение не улучшает)

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
LEARNING RATE, Lr	Шаг обновления параметров в градиентном спуске; гиперпараметр, определяющий, насколько сильно изменяются веса на каждой итерации. Слишком высокий learning rate вызывает расходимость (<i>oscillation</i>), слишком низкий – медленное обучение или застревание в локальном минимуме	Тестирование <code>learning_rate</code> : $lr = 0,1$ – loss растёт, модель расходится; $lr = 0,01$ – быстрое снижение loss, сходимость за 50 эпох; $lr = 0,001$ – очень медленное снижение, после 100 эпох loss все еще высокий; $lr = 0,0001$ – обучение практически не идет. Оптимально $lr = 0,01$
ACTIVATION FUNCTION	Функция, применяемая к взвешенной сумме входов нейрона, вводящая нелинейность в нейронную сеть. Без нелинейных функций активации многослойная сеть была бы эквивалентна одному линейному слою. Популярные функции: ReLU, Sigmoid, Tanh, Softmax	Нейрон вычислил, что $z = 2,5$ (взвешенная сумма входов). Применяется ReLU: $output = \max(0, 2,5) = 2,5$. Для другого нейрона $z = -1,2$; ReLU: $output = \max(0, -1,2) = 0$ (нейрон «мертв», не активируется)
RELU (RECTIFIED LINEAR UNIT)	Линейная функция активации с выпрямлением: $f(x) = \max(0, x)$. ReLU возвращает вход, если он положительный, и 0 в противном случае. ReLU – самая популярная функция активации в глубоком обучении благодаря простоте вычисления и решению проблемы затухающего градиента	Сверточная сеть для распознавания изображений использует ReLU после каждого сверточного слоя: если нейрон выдал $-3,2$, ReLU обнуляет (0); если выдал $5,7$, ReLU пропускает (5,7). ReLU ускоряет обучение в 6 раз по сравнению с sigmoid и tanh
SOFTMAX	Функция активации для многоклассовой классификации, преобразующая вектор сырых оценок (logits) в распределение вероятностей, где	Нейросеть классифицирует изображение на 3 класса (кошка, собака, птица). Выходной слой выдал logits: $[2,1; 0,5; -0,3]$. Softmax

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	все значения положительны и суммируются в 1. Используется в выходном слое для задач с $K > 2$ классов	преобразует: $P(\text{кошка}) = 0,78$; $P(\text{собака}) = 0,16$; $P(\text{птица}) = 0,06$. Сумма вероятностей = 1,0. Предсказание: кошка (наивысшая вероятность)
CONVOLUTIONAL NEURAL NETWORKS, CNN	Специализированные нейронные сети для обработки изображений и других данных с сеточной структурой. CNN используют сверточные слои (<i>convolution</i>), которые применяют фильтры (ядра) к локальным областям входных данных, автоматически извлекая пространственные признаки	Распознавание рукописных цифр MNIST: CNN принимает изображение 28×28 пикселей. Первый сверточный слой применяет 32 фильтра 3×3 , извлекая простые признаки (края, углы). Второй слой – 64 фильтра, извлекает более сложные паттерны. После pooling и fully connected слоев сеть классифицирует цифру с точностью 99,2 %
RECURRENT NEURAL NETWORKS, RNN	Нейронные сети для обработки последовательных данных (временные ряды, текст, речь), где выход на предыдущем шаге передается как вход на следующем шаге. RNN имеют «память» о предыдущих элементах последовательности	Прогнозирование курса акций: RNN получает на вход цены за последние 30 дней. На шаге $t = 1$ обрабатывает цену дня 1, сохраняет скрытое состояние h_1 . На шаге $t = 2$ обрабатывает цену дня 2 вместе с h_1 , получает h_2 , и так далее до дня 30. Финальное скрытое состояние h_{30} содержит информацию обо всей последовательности для прогноза на день 31
TRANSFORMER	Архитектура нейронной сети на основе механизма внимания (<i>attention mechanism</i>), революционизировавшая об-	Модель машинного перевода English → Russian на базе Transformer (как в Яндекс Переводчике):

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	работку естественного языка. Обрабатывает всю последовательность параллельно (в отличие от RNN) и эффективно улавливает долгосрочные зависимости	входное предложение “I love machine learning” разбивается на токены, каждый токен преобразуется в вектор. Self-attention вычисляет взаимосвязи между всеми словами параллельно. Encoder кодирует смысл предложения, Decoder генерирует перевод «Я люблю машинное обучение» токен за токеном
ATTENTION MECHANISM	Механизм, позволяющий модели фокусироваться на релевантных частях входных данных при обработке каждого элемента. Attention вычисляет веса важности для каждого элемента последовательности, позволяя модели «обращать внимание» на нужные слова/признаки	Перевод предложения “The cat sat on the mat because it was tired”. При переводе слова “it” механизм внимания назначает высокие веса словам “cat” (0,7) и “tired” (0,2), низкие – остальным. Это помогает модели понять, что “it” относится к коту, а не к коврику, и правильно перевести
TRANSFER LEARNING	Использование предварительно обученной модели (<i>pre-trained model</i>) для решения новой, но связанной задачи. Вместо обучения с нуля берется модель, обученная на большой базе данных (например, ImageNet), и дообучается (<i>fine-tuning</i>) на меньшем наборе данных конкретной задачи	Классификация рентгеновских снимков на пневмонию/норму (всего 5000 снимков). Вместо обучения CNN с нуля берется предобученная ResNet-50 (обучена на ImageNet, 1,2 млн изображений). Последние слои заменяются и дообучаются на медицинских снимках. Точность 94 % против 78 % при обучении с нуля – трансферное обучение дало +16 %

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
ENSEMBLE METHODS	Комбинация нескольких моделей (ансамбль) для получения лучших результатов, чем у отдельных моделей. Ансамбли снижают дисперсию, смещение или и то, и другое. Основные подходы: бэггинг (<i>bagging</i>), бустинг (<i>boosting</i>), стэкинг (<i>stacking</i>)	Прогнозирование оттока клиентов: отдельное дерево решений дает точность 76 %, случайный лес (ансамбль из 100 деревьев, бэггинг) – 84 %, градиентный бустинг (ансамбль из 200 слабых деревьев, бустинг) – 87 %. Ансамбли значительно превосходят одиночные модели
BAGGING	Метод ансамбля, обучающий множество моделей параллельно на различных бутстрэп-выборках (случайных подвыборках с возвращением) из тренировочных данных. Итоговое предсказание – усреднение (регрессия) или голосование (классификация). Bagging снижает дисперсию	Random Forest (случайный лес) – классический пример бэггинга: обучается 100 решающих деревьев. Каждое дерево обучается на своей бутстрэп-выборке (случайные 80 % данных с повторениями). При предсказании все 100 деревьев голосуют, итоговый класс – большинство голосов. Точность 89 % против 76 % у 1 дерева
BOOSTING	Метод ансамбля с последовательным обучением, где каждая следующая модель фокусируется на ошибках предыдущих. Слабые модели (<i>weak learners</i>) объединяются в сильную (<i>strong learner</i>). Популярные алгоритмы: AdaBoost, Gradient Boosting, XGBoost, LightGBM	Классификация спама: первое дерево правильно классифицирует 70 % писем. Второе дерево обучается на ошибках первого (больше веса неправильно классифицированным письмам), точность на сложных случаях 65 %. Третье дерево исправляет ошибки первых двух. После 100 деревьев ансамбль достигает точности 96 %

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
RANDOM FOREST	Ансамбль решающих деревьев, сочетающий бэггинг (случайные выборки данных) и случайный выбор признаков в каждом узле дерева. Random Forest снижает переобучение отдельных деревьев и обеспечивает высокую точность	Предсказание стоимости аренды квартир: Random Forest из 200 деревьев учитывает 15 признаков (площадь, район, этаж, метро и т.д.). Каждое дерево строится на случайной подвыборке данных, в каждом узле выбирается случайное подмножество признаков ($\sqrt{15} \approx 4$ признака). Итоговая точность $R^2 = 0,87$
SCALABILITY	Способность системы справляться с ростом объема данных, числа пользователей или вычислительной сложности без существенной потери производительности. Масштабируемость критична для развертывания ML-моделей в промышленную эксплуатацию	Модель рекомендаций YouTube должна обрабатывать 1 млрд запросов в день от 2 млрд пользователей. Система масштабируется горизонтально: добавление новых серверов с 10 до 100 позволяет обрабатывать в 10 раз больше запросов с линейным ростом производительности
DISTRIBUTED COMPUTING	Обработка данных на нескольких узлах (компьютерах, серверах) параллельно для ускорения вычислений и обработки больших объемов данных, которые не помещаются на одном компьютере	Обучение модели на 10 ТБ логов пользователей: один сервер с 1 ТБ диска не справится. Распределенная система из 50 узлов (каждый с 1 ТБ) разделяет данные: каждый узел обрабатывает свои 200 Гб параллельно. Обучение завершается за 4 ч вместо 200 ч на одном сервере
HADOOP	Фреймворк для распределенной обработки больших данных с открытым исходным	Анализ логов веб-сайта (50 ТБ за месяц): данные хранятся в HDFS на 100 узлах.

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	кодом, использующий модель программирования MapReduce. Hadoop состоит из HDFS (распределенная файловая система) и MapReduce (параллельная обработка)	MapReduce-задача: Map – подсчет просмотров каждой страницы на каждом узле; Reduce – агрегация результатов со всех узлов. Итоговый отчет по 10 000 страниц генерируется за 30 мин
SPARK	Быстрый движок для обработки больших данных, поддерживающий итеративные алгоритмы и интерактивные запросы. Хранит данные в памяти (в отличие от Hadoop MapReduce, который пишет на диск), что делает его в 10–100 раз быстрее для многих задач	Обучение модели машинного обучения на 1 ТБ данных: Hadoop MapReduce требует 10 итераций, каждая читает данные с диска 5 ч. Apache Spark загружает данные в память один раз, 10 итераций работают с памятью 15 мин. Ускорение в 20 раз
ONLINE LEARNING	Обучение модели инкрементально по мере поступления новых данных, без необходимости переобучать на всех данных заново. Онлайн обучение адаптируется к изменениям в данных (<i>concept drift</i>) и работает с потоковыми данными	Модель предсказания кликов на рекламу (CTR): каждые 5 мин приходят новые данные о 10 000 показов и кликов. Онлайн-алгоритм (SGD) обновляет веса модели на новых данных, адаптируясь к изменениям поведения пользователей. Модель всегда актуальна без полного переобучения
REINFORCEMENT LEARNING, RL	Тип машинного обучения, при котором агент обучается через взаимодействие со средой: выполняет действия, получает награды (<i>reward</i>) или штрафы и учится максимизировать	Обучение агента играть в шахматы: агент делает ход (действие), среда (игра) переходит в новое состояние (позиция на доске), агент получает награду (+1 за победу; –1 за поражение;

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	совокупную награду. RL используется в играх, робототехнике, автономных системах	0 за ничью). После миллионов игр агент учится стратегии, побеждающей гроссмейстеров (AlphaZero)
GENERATIVE AI	Искусственный интеллект, способный создавать новый контент (текст, изображения, музыку, код), похожий на созданный человеком. Генеративные модели (GPT, DALL-E, Stable Diffusion) обучаются на больших данных и генерируют новые примеры из выученного распределения	ChatGPT генерирует ответ на вопрос пользователя: модель анализирует контекст диалога, предсказывает следующее слово, затем следующее, создавая связный осмысленный текст. Midjourney создает изображение по описанию «космонавт верхом на лошади в стиле киберпанка» – изображение, которого не существовало до генерации
LARGE LANGUAGE MODELS, LLM	Модели для обработки естественного языка с миллиардами параметров, обученные на огромных текстовых корпусах. LLM (GPT-4, PaLM, LLaMA) способны генерировать текст, отвечать на вопросы, переводить, писать код и выполнять множество задач без специального дообучения	GPT-4 с 1,7 трлн параметров обучена на сотнях миллиардов слов из Интернета, книг, статей. Модель может написать эссе о влиянии ИИ на экономику, решить математическую задачу с объяснением, перевести текст с китайского на суахили, отладить код на Python – все без дополнительного обучения
ETHICAL AI	Разработка и использование искусственного интеллекта с учетом этических принципов: справедливости (<i>fairness</i>), прозрачности (<i>transparency</i>), подотчетности (<i>accountability</i>), конфиденциальности (<i>privacy</i>), безопасности.	Система кредитного скоринга проверяется на этичность: модель не должна дискриминировать по полу, возрасту, расе (<i>fairness</i>). Решения объяснимы – можно понять, почему заявка отклонена (<i>transparency</i>). При ошибке можно

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	Этичный ИИ предотвращает вред и дискриминацию	оспорить решение (<i>accountability</i>). Данные клиентов защищены (<i>privacy</i>)
FAIRNESS	Отсутствие дискриминации в алгоритмах; требование, чтобы ИИ-системы не воспроизводили и не усиливали предвзятость (<i>biases</i>) против определенных групп людей (по полу, расе, возрасту и т.д.)	Алгоритм найма анализирует резюме: тестирование показывает, что модель отклоняет женщин, претендующих на технические позиции, в 2 раза чаще мужчин с аналогичной квалификацией (<i>bias</i>). После корректировки (удаление признаков пола, <i>rebalancing</i> данных) разница сокращается до 5 % – модель стала справедливее
EXPLAINABILITY	Возможность понять и интерпретировать решения модели; способность объяснить, почему модель приняла то или иное решение. Explainability критична для критических приложений (медицина, финансы, юриспруденция) и построения доверия к ИИ	Модель отказала в кредите. Система объяснимости (SHAP) демонстрирует, что основные причины отказа – высокая долговая нагрузка (вклад 45 %), короткая кредитная история (вклад 30 %), низкий доход (вклад 25 %). Клиент понимает причины и может исправить ситуацию (увеличить доход, погасить долги)
INTERPRETABILITY	Степень понимания причинно-следственных связей в модели, т.е. насколько человек может понять логику принятия решений моделью. Интерпретируемые модели (линейная регрессия, деревья решений) прозрачнее	Линейная регрессия предсказывает цену квартиры: $\text{цена} = 50\,000 \times \text{площадь} + 500\,000 \times (\text{близость к метро}) - 200\,000 \times \text{этаж}$. Интерпретация очевидна: каждый квадратный метр добавляет 50 тыс. руб., близость

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	«черных ящиков» (глубокие нейросети)	к метро добавляет 500 тыс. руб., каждый этаж выше первого уменьшает цену на 200 тыс. руб.
MODEL DRIFT	Изменение производительности модели со временем из-за изменения статистических свойств данных или целевой переменной. Model drift требует мониторинга и переобучения моделей	Модель предсказания спроса на товары обучена на данных 2019 г. В 2020 г. началась пандемия, поведение покупателей изменилось: спрос на маски и антисептики вырос в 100 раз, на одежду для офиса упал. Модель, не переобученная на новых данных, дает огромные ошибки прогноза – произошел model drift
MLOPS	Практика развертывания и поддержки моделей машинного обучения в продакшене, объединяющая машинное обучение, DevOps и data engineering. MLOps включает автоматизацию обучения, мониторинг моделей, версионирование, CI/CD для ML	Компания развернула 50 ML-моделей в продакшене. MLOps-платформа автоматически отслеживает качество моделей (мониторинг drift), переобучает модели на новых данных каждую неделю (автоматизация), тестирует новые версии перед развертыванием (CI/CD), откатывает модель при падении качества (безопасность)
FEATURE ENGINEERING	Процесс создания, отбора и преобразования признаков (<i>features</i>) из сырых данных для улучшения производительности моделей машинного обучения. Feature Engineering часто влияет	Исходные данные: дата транзакции, сумма, магазин. <i>Feature Engineering</i> создает новые признаки: день недели (понедельник = 0, ..., воскресенье = 6); время суток

TERM (ENGLISH)	RUSSIAN DEFINITION	EXAMPLE
	на результат сильнее, чем выбор алгоритма	(утро/день/вечер/ночь); день месяца (начало/середина/конец); сумма относительно среднего чека клиента (аномально большая/маленькая); время, прошедшее с последней покупки. Новые признаки повышают точность модели детекции мошенничества с 78 до 91 %
AUTOML	Автоматизация процесса машинного обучения: автоматический выбор модели, подбор гиперпараметров, конструирование признаков, валидация. AutoML позволяет non-experts создавать эффективные ML-модели без глубоких знаний	Бизнес-аналитик без опыта в ML загружает данные о продажах (100 000 строк, 20 признаков) в AutoML платформу (Google AutoML, H2O.ai). Система автоматически пробует 15 алгоритмов (от линейной регрессии до градиентного бустинга), подбирает гиперпараметры для каждого, выбирает лучшую модель (XGBoost с точностью 92 %), создает API для предсказаний. Весь процесс занял 2 ч ручного вмешательства

Index of Terms

TERM	SECTION COLOUR	SECTION TITLE
ACCURACY		Section 3. Statistical models for supervised learning
ACTIVATION FUNCTION		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ADJUSTED R -SQUARED		Section 3. Statistical models for supervised learning
ALTERNATIVE HYPOTHESIS, H_1		Section 2. Exploratory data analysis and descriptive statistics
ANOMALY DETECTION		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ANOVA		Section 2. Exploratory data analysis and descriptive statistics
ARTIFICIAL INTELLIGENCE		Section 1. Fundamentals of probability theory and statistical thinking
ASSOCIATION RULES		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ATTENTION MECHANISM		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
AUTOML		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
AUC		Section 3. Statistical models for supervised learning
BACKPROPAGATION		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
BAGGING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
BAYESIAN CLASSIFICATION		Section 3. Statistical models for supervised learning
BAYES THEOREM		Section 1. Fundamentals of probability theory and statistical thinking
BATCH SIZE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration

TERM	SECTION COLOUR	SECTION TITLE
BERNOULLI DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
BIG DATA		Section 1. Fundamentals of probability theory and statistical thinking
BINARY CLASSIFICATION		Section 3. Statistical models for supervised learning
BINOMIAL DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
BIAS-VARIANCE TRADEOFF		Section 3. Statistical models for supervised learning
BOX PLOT		Section 2. Exploratory data analysis and descriptive statistics
BOOSTING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
CENTROID		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
CHI-SQUARE TEST		Section 2. Exploratory data analysis and descriptive statistics
CLASSIFICATION		Section 3. Statistical models for supervised learning
CLUSTERING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
CONFIDENCE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
CONFIDENCE INTERVAL		Section 2. Exploratory data analysis and descriptive statistics
CONFIDENCE LEVEL		Section 2. Exploratory data analysis and descriptive statistics
CENTRAL LIMIT THEOREM		Section 1. Fundamentals of probability theory and statistical thinking
CONDITIONAL PROBABILITY		Section 1. Fundamentals of probability theory and statistical thinking
CONFUSION MATRIX		Section 3. Statistical models for supervised learning
CONTINUOUS RANDOM VARIABLE		Section 1. Fundamentals of probability theory and statistical thinking
CONVOLUTIONAL NEURAL NETWORKS, CNN		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration

TERM	SECTION COLOUR	SECTION TITLE
CORRELATION		Section 2. Exploratory data analysis and descriptive statistics
COVARIANCE		Section 2. Exploratory data analysis and descriptive statistics
CRITICAL REGION		Section 2. Exploratory data analysis and descriptive statistics
CROSS-VALIDATION		Section 3. Statistical models for supervised learning
CUMULATIVE DISTRIBUTION FUNCTION, CDF		Section 1. Fundamentals of probability theory and statistical thinking
DATA CLEANING		Section 2. Exploratory data analysis and descriptive statistics
DATA MINING		Section 2. Exploratory data analysis and descriptive statistics
DATA SCIENCE		Section 2. Exploratory data analysis and descriptive statistics
DATA VISUALIZATION		Section 2. Exploratory data analysis and descriptive statistics
DBSCAN		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
DECISION BOUNDARY		Section 3. Statistical models for supervised learning
DEEP LEARNING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
DENDROGRAM		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
DEPENDENT VARIABLE		Section 3. Statistical models for supervised learning
DESCRIPTIVE STATISTICS		Section 2. Exploratory data analysis and descriptive statistics
DIMENSIONALITY REDUCTION		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
DISCRETE RANDOM VARIABLE		Section 1. Fundamentals of probability theory and statistical thinking
DISTRIBUTED COMPUTING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
EIGENVALUE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration

TERM	SECTION COLOUR	SECTION TITLE
EIGENVECTOR		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ELBOW METHOD		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ELASTIC NET		Section 3. Statistical models for supervised learning
EPOCH		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ENSEMBLE METHODS		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
ETHICAL AI		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
EVENT		Section 1. Fundamentals of probability theory and statistical thinking
EXPECTATION-MAXIMIZATION, EM		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
EXPLAINED VARIANCE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
EXPLORATORY DATA ANALYSIS, EDA		Section 2. Exploratory data analysis and descriptive statistics
EXPECTED VALUE		Section 1. Fundamentals of probability theory and statistical thinking
EXPLAINABILITY		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
F -TEST		Section 2. Exploratory data analysis and descriptive statistics
$F1$ -SCORE		Section 3. Statistical models for supervised learning
FAIRNESS		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
FALSE NEGATIVE, FN		Section 3. Statistical models for supervised learning
FALSE POSITIVE, FP		Section 3. Statistical models for supervised learning

TERM	SECTION COLOUR	SECTION TITLE
FALSE POSITIVE RATE, FPR		Section 3. Statistical models for supervised learning
FEATURE		Section 3. Statistical models for supervised learning
FEATURE ENGINEERING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
GAUSSIAN MIXTURE MODELS, GMM		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
GENERATIVE AI		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
GRADIENT DESCENT		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
GRID SEARCH		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
HADOOP		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
HIERARCHICAL CLUSTERING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
HISTOGRAM		Section 2. Exploratory data analysis and descriptive statistics
HOMOSCEDASTICITY		Section 3. Statistical models for supervised learning
HETEROSCEDASTICITY		Section 3. Statistical models for supervised learning
HYPERPARAMETERS		Section 3. Statistical models for supervised learning
HYPOTHESIS TESTING		Section 2. Exploratory data analysis and descriptive statistics
INDEPENDENCE		Section 1. Fundamentals of probability theory and statistical thinking
INDEPENDENT VARIABLE		Section 3. Statistical models for supervised learning
INTERCEPT		Section 3. Statistical models for supervised learning
INTERQUARTILE RANGE, IQR		Section 2. Exploratory data analysis and descriptive statistics

TERM	SECTION COLOUR	SECTION TITLE
INTERPRETABILITY		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
INTERVAL ESTIMATION		Section 2. Exploratory data analysis and descriptive statistics
JOINT PROBABILITY		Section 1. Fundamentals of probability theory and statistical thinking
K-MEANS CLUSTERING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
K-FOLD CROSS-VALIDATION		Section 3. Statistical models for supervised learning
KURTOSIS		Section 2. Exploratory data analysis and descriptive statistics
LARGE LANGUAGE MODELS, LLM		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
LASSO REGRESSION		Section 3. Statistical models for supervised learning
LAW OF LARGE NUMBERS		Section 1. Fundamentals of probability theory and statistical thinking
LEARNING RATE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
LEAVE-ONE-OUT, LOO		Section 3. Statistical models for supervised learning
LIFT		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
LIKELIHOOD		Section 3. Statistical models for supervised learning
LINEAR REGRESSION		Section 3. Statistical models for supervised learning.
LOGISTIC REGRESSION		Section 3. Statistical models for supervised learning
MACHINE LEARNING		Section 1. Fundamentals of probability theory and statistical thinking
MARGINAL PROBABILITY		Section 1. Fundamentals of probability theory and statistical thinking
MEAN		Section 2. Exploratory data analysis and descriptive statistics
MEAN ABSOLUTE ERROR, MAE		Section 3. Statistical models for supervised learning

TERM	SECTION COLOUR	SECTION TITLE
MEAN SQUARED ERROR, MSE		Section 3. Statistical models for supervised learning
MEASURES OF CENTRAL TENDENCY		Section 2. Exploratory data analysis and descriptive statistics
MEASURES OF DISPERSION		Section 2. Exploratory data analysis and descriptive statistics
MEDIAN		Section 2. Exploratory data analysis and descriptive statistics
MISSING DATA		Section 2. Exploratory data analysis and descriptive statistics
MLOPS		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
MODE		Section 2. Exploratory data analysis and descriptive statistics
MODEL DRIFT		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
MULTICOLLINEARITY		Section 3. Statistical models for supervised learning
MULTIPLE LINEAR REGRESSION		Section 3. Statistical models for supervised learning
MULTICLASS CLASSIFICATION		Section 3. Statistical models for supervised learning
NAIVE BAYES		Section 3. Statistical models for supervised learning
NEURAL NETWORK		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
NORMAL DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
NULL HYPOTHESIS, H_0		Section 2. Exploratory data analysis and descriptive statistics
ONLINE LEARNING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
OUTLIER		Section 2. Exploratory data analysis and descriptive statistics
OVERFITTING		Section 3. Statistical models for supervised learning
P-VALUE		Section 2. Exploratory data analysis and descriptive statistics

TERM	SECTION COLOUR	SECTION TITLE
PEARSON CORRELATION		Section 2. Exploratory data analysis and descriptive statistics
PERCENTILES		Section 2. Exploratory data analysis and descriptive statistics
POINT ESTIMATION		Section 2. Exploratory data analysis and descriptive statistics
POISSON DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
POLYNOMIAL REGRESSION		Section 3. Statistical models for supervised learning
POSTERIOR PROBABILITY		Section 3. Statistical models for supervised learning
POWER OF TEST		Section 2. Exploratory data analysis and descriptive statistics
PRECISION		Section 3. Statistical models for supervised learning
PREDICTIVE ANALYTICS		Section 1. Fundamentals of probability theory and statistical thinking
PRIOR PROBABILITY		Section 3. Statistical models for supervised learning
PROBABILITY DENSITY FUNCTION, PDF		Section 1. Fundamentals of probability theory and statistical thinking
PROBABILITY DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
PROBABILITY MASS FUNCTION, PMF		Section 1. Fundamentals of probability theory and statistical thinking
PROBABILITY THRESHOLD		Section 3. Statistical models for supervised learning
PRINCIPAL COMPONENT ANALYSIS, PCA		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
QUANTILES		Section 2. Exploratory data analysis and descriptive statistics
RANDOM EXPERIMENT		Section 1. Fundamentals of probability theory and statistical thinking
RANDOM FOREST		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
RANDOM VARIABLE		Section 1. Fundamentals of probability theory and statistical thinking
RANGE		Section 2. Exploratory data analysis and descriptive statistics
RECALL		Section 3. Statistical models for supervised learning

TERM	SECTION COLOUR	SECTION TITLE
RECURRENT NEURAL NETWORKS, RNN		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
REGRESSION		Section 3. Statistical models for supervised learning
REGRESSION COEFFICIENTS		Section 3. Statistical models for supervised learning
REGULARIZATION		Section 3. Statistical models for supervised learning
RELU		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
REINFORCEMENT LEARNING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
RESIDUAL		Section 3. Statistical models for supervised learning
RIDGE REGRESSION		Section 3. Statistical models for supervised learning
ROOT MEAN SQUARED ERROR, RMSE		Section 3. Statistical models for supervised learning
ROC CURVE		Section 3. Statistical models for supervised learning
SAMPLE SIZE		Section 3. Statistical models for supervised learning
SAMPLE SPACE		Section 1. Fundamentals of probability theory and statistical thinking
SCATTER PLOT		Section 2. Exploratory data analysis and descriptive statistics
SCALABILITY		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
SENSITIVITY		Section 3. Statistical models for supervised learning
SIGMOID FUNCTION		Section 3. Statistical models for supervised learning
SIGNIFICANCE LEVEL, α		Section 2. Exploratory data analysis and descriptive statistics
SILHOUETTE SCORE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
SIMPLE LINEAR REGRESSION		Section 3. Statistical models for supervised learning

TERM	SECTION COLOUR	SECTION TITLE
SKEWNESS		Section 2. Exploratory data analysis and descriptive statistics
SOFTMAX		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
SPARK		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
SPEARMAN CORRELATION		Section 2. Exploratory data analysis and descriptive statistics
SPECIFICITY		Section 3. Statistical models for supervised learning
STANDARD DEVIATION		Section 1. Fundamentals of probability theory and statistical thinking
STATISTICAL INFERENCE		Section 2. Exploratory data analysis and descriptive statistics
STATISTICAL LEARNING		Section 1. Fundamentals of probability theory and statistical thinking
STOCHASTIC GRADIENT DESCENT, SGD		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
SUM OF SQUARED ERRORS, SSE		Section 3. Statistical models for supervised learning
SUPERVISED LEARNING		Section 3. Statistical models for supervised learning
SUPPOR		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
TARGET VARIABLE		Section 3. Statistical models for supervised learning
TEST SET		Section 3. Statistical models for supervised learning
TEST STATISTIC		Section 2. Exploratory data analysis and descriptive statistics
<i>T</i> -TEST		Section 2. Exploratory data analysis and descriptive statistics
<i>T</i> -SNE		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
TRAINING SET		Section 3. Statistical models for supervised learning
TRANSFER LEARNING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration

TERM	SECTION COLOUR	SECTION TITLE
TRANSFORMER		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
TRUE NEGATIVE, TN		Section 3. Statistical models for supervised learning
TRUE POSITIVE, TP		Section 3. Statistical models for supervised learning
TRUE POSITIVE RATE, TPR		Section 3. Statistical models for supervised learning
TYPE I ERROR		Section 2. Exploratory data analysis and descriptive statistics
TYPE II ERROR		Section 2. Exploratory data analysis and descriptive statistics
UMAP		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
UNDERFITTING		Section 3. Statistical models for supervised learning
UNIFORM DISTRIBUTION		Section 1. Fundamentals of probability theory and statistical thinking
UNSUPERVISED LEARNING		Section 4. Advanced topics: unsupervised techniques, scalability, and AI integration
VALIDATION SET		Section 3. Statistical models for supervised learning
VARIANCE		Section 1. Fundamentals of probability theory and statistical thinking
VARIETY		Section 1. Fundamentals of probability theory and statistical thinking
VELOCITY		Section 1. Fundamentals of probability theory and statistical thinking
VOLUME		Section 1. Fundamentals of probability theory and statistical thinking
Z-TEST		Section 2. Exploratory data analysis and descriptive statistics

РАЗДЕЛ 1

ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И СТАТИСТИЧЕСКОГО МЫШЛЕНИЯ

1.1. Введение в большие данные и роль статистики в искусственном интеллекте

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
БОЛЬШИЕ ДАННЫЕ (BIG DATA)	Big Data refers to extremely large and complex datasets that require specialized methods and technologies for processing and analysis. In the context of artificial intelligence, big data serves as a fundamental resource for training machine learning models, enabling algorithms to identify complex patterns and dependencies	The Safe City facial recognition system processes millions of images from surveillance cameras to train a neural network that automatically identifies people in video streams
ОБЪЕМ ДАННЫХ	Volume is a characteristic describing the scale of data. In AI tasks, data volume directly impacts model quality: the more labeled data available, the more accurately the algorithm can learn. However, this also requires greater computational resources and memory	To train the Yandex. Translate machine translation model, billions of sentence pairs in different languages are used, requiring petabytes of data storage
СКОРОСТЬ ГЕНЕРАЦИИ ДАННЫХ	Velocity is a characteristic describing the speed of data generation and processing. In real-time systems such as autonomous vehicles or high-frequency trading algorithms, data must be processed within milliseconds to enable instant decision-making	The Yandex.Music recommendation system analyzes user actions (plays, skips, likes) in real-time and instantly adapts the playlist to current preferences

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
РАЗНООБРАЗИЕ ДАННЫХ	Variety is a characteristic describing various formats and types of data. Modern AI systems work with multi-modal data: text, images, video, audio, numerical indicators, geospatial data, and time series	The YandexGPT multimodal model simultaneously analyzes text queries, images, and voice commands to provide comprehensive answers to users
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ	Artificial Intelligence is a field of computer science focused on creating systems capable of performing tasks requiring human intelligence: pattern recognition, natural language understanding, decision-making, learning from experience, and solving complex problems	The Alisa voice assistant uses artificial intelligence for speech recognition, understanding user intent, searching for information, and conducting conversations in natural language
МАШИННОЕ ОБУЧЕНИЕ	Machine Learning is a subset of artificial intelligence that uses statistical methods to automatically improve performance based on data. Machine learning algorithms build mathematical models based on training data to make predictions or decisions without explicit programming	Sberbank's machine learning algorithm analyzes the history of credit applications and customer payments to predict the probability of default for new borrowers
НАУКА О ДАННЫХ	Data Science is an interdisciplinary field using scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data Science combines statistics, machine learning, data analysis, and domain expertise	A data scientist at Wildberries analyzes customer behavior, demand seasonality, and external factors to optimize warehouse inventory and forecast sales
СТАТИСТИЧЕСКОЕ ОБУЧЕНИЕ	Statistical Learning is a set of tools for modeling and understanding complex datasets using statistical methods. It	A statistical learning method (linear regression) is used on the Cian service to predict

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	includes regression, classification, clustering, and dimensionality reduction methods to identify patterns in data	apartment prices based on their characteristics: area, floor, district, and year of construction
ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ	Data Mining is the process of discovering hidden patterns, anomalies, correlations, and knowledge in large volumes of data using methods at the intersection of statistics, machine learning, and database analysis	The Pyaterochka chain stores apply Data Mining to analyze customer receipts and discover association rules: “customers who buy beer often buy chips”, which optimizes product placement
ПРЕДИКТИВНАЯ АНАЛИТИКА	Predictive Analytics is the use of statistical algorithms and machine learning methods to forecast future events based on historical data and identify probabilistic trends	Predictive analytics at Aeroflot airline forecasts demand for tickets on different routes, enabling dynamic price changes and flight load optimization

1.2. Теория вероятностей для науки о данных

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
СЛУЧАЙНАЯ ВЕЛИЧИНА	A random variable is a quantity that takes one possible value as a result of an experiment, unknown in advance and dependent on random causes. In machine learning, random variables model uncertainty in data and predictions	The number of user clicks on an advertising banner per day is a random variable used to optimize advertising campaigns in “Yandex Direct”
ДИСКРЕТНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА	A discrete random variable is a random variable that can take a finite or countable set of values. Discrete variables are characteristic of classification tasks, event counting, and categorical data	The number of errors in code before successful compilation is a discrete random variable that can be used to assess development quality

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
НЕПРЕРЫВНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА	A continuous random variable is a random variable that can take any values from a certain interval (finite or infinite). Continuous variables are used to model physical measurements, time, and probabilities	Server response time (in milliseconds) is a continuous random variable critically important for monitoring web-service performance
РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ	A probability distribution is a function or rule describing the probability of a random variable taking various possible values. Probability distributions are the foundation of statistical modeling in AI	The delivery time distribution for “Yandex Food” orders is described by a lognormal distribution: most deliveries take 30–45 minutes, but longer times are possible
ФУНКЦИЯ ВЕРОЯТНОСТИ	Probability Mass Function is a function that defines the probability distribution for a discrete random variable. PMF assigns a probability to each possible value, with the sum of all probabilities equal to 1	Probability function for the number of support requests per hour: $P(0) = 0,1$; $P(1) = 0,2$; $P(2) = 0,3$; $P(3) = 0,25$; $P(4) = 0,15$
ФУНКЦИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ	Probability Density Function is a function describing the probability density for a continuous random variable. The area under the PDF curve on interval $[a, b]$ equals the probability that the random variable falls within that interval	The battery life density distribution of a smartphone describes that 8–10 hours of operation is most likely, while values less than 4 or more than 14 hours are unlikely
ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ	Cumulative Distribution Function is a function defining the probability that a random variable takes a value less than or equal to a given value. CDF is used to calculate percentiles and quantiles in data analysis	The page load time distribution function shows that 90 % of users see the page fully loaded within 2 seconds or faster (Page Load Time metric)
МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ	Expected Value is the weighted average value of a random variable, where	The expected value of product sales per day is 15 units – this value

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	weights are probabilities. Expected value characterizes the “center” of the distribution and is used as a forecast in regression tasks	is used to optimize warehouse inventory
ДИСПЕРСИЯ	Variance is a measure of the spread of a random variable around its expected value. Variance shows the degree of uncertainty or risk: high variance means greater variability of results	The variance of an investment portfolio’s return shows risk: two portfolios with the same expected return of 10 % can have different variances – 2 % (stable) or 15 % (risky)
СТАНДАРТНОЕ ОТКЛОНЕНИЕ	The square root of variance, measured in the same units as the random variable itself. Standard deviation is interpreted as the “typical deviation” from the mean and is used to assess model accuracy. In machine learning, standard deviation helps evaluate the stability of predictions: a low standard deviation indicates that data points are clustered close to the mean (consistent model performance), while a high standard deviation indicates that values are spread out over a wider range (high variability or uncertainty)	The model’s prediction error has a standard deviation of 5 %: if the average accuracy is 85 %, most predictions fall in the range of 80–90 % accuracy. The standard deviation of food delivery time is 8 minutes: if the average time is 35 minutes, most deliveries take between 27 and 43 minutes
УСЛОВНАЯ ВЕРОЯТНОСТЬ	Conditional Probability is the probability of an event given that another event has already occurred. Conditional probabilities underlie Bayesian networks, naive Bayesian classifiers, and many AI algorithms	The probability that a user will buy a product (event <i>A</i>), given that they added it to the cart (event <i>B</i>), is 0,65 – this is used to predict conversion in online stores
ТЕОРЕМА БАЙЕСА	Bayes Theorem is a fundamental theorem of probability theory that quantitatively	A fraud detection system uses Bayes Theorem: if a transaction is made

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	determines the probability of an event based on prior knowledge of conditions that may be related to the event. Bayes Theorem is the basis of Bayesian machine learning and spam filtering	in an unusual location and for a large amount, the probability of fraud increases from 0,1 % (prior) to 78 % (posterior)
НЕЗАВИСИМОСТЬ СОБЫТИЙ	Independence is a property of events where the occurrence of one does not affect the probability of the other. The assumption of independence simplifies calculations in probabilistic models, but real data often shows dependence	Dice roll results at different times are independent – this property is used when generating random numbers for initializing neural network weights
СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ	A random experiment is a process whose result is unpredictable, but the set of all possible outcomes is known. In machine learning, random experiments model the stochasticity of data and algorithms	<i>A/B</i> testing of two landing page versions is a random experiment where the result (conversion) depends on random distribution of users between variants
ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ ИСХОДОВ	Sample Space is the set of all possible outcomes of a random experiment. The outcome space defines the range of values in which the probabilistic model operates	For a binary classification task (spam/not spam), the outcome space consists of two elements: {spam, not spam}. For MNIST digit recognition – {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
СОБЫТИЕ	An event is a subset of the sample space with a certain property. Events are used to formulate hypotheses in statistical testing and <i>A/B</i> experiments	The event “a user made a purchase” includes outcomes: purchase of any amount, in any product category, by any payment method
СОВМЕСТНАЯ ВЕРОЯТНОСТЬ	Joint Probability is the probability of two or more events occurring simultaneously. Joint probability distributions model relationships between features in multidimensional data	The joint probability that a user will both click on an ad AND buy a product is 0,03 – this is used to calculate expected profit from an advertising campaign

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
МАРГИНАЛЬНАЯ ВЕРОЯТНОСТЬ	Marginal Probability is the probability of an event without considering other events, obtained by summing (or integrating) the joint probability over all possible values of other variables	The probability of purchasing a product (regardless of whether the user clicked on an ad or came organically) is 0,12 – this is the marginal conversion probability
РАСПРЕДЕЛЕНИЕ БЕРНУЛЛИ	Bernoulli Distribution is a discrete distribution with two possible outcomes: success (1) with probability p and failure (0) with probability $1 - p$. Bernoulli distribution models binary events in classification	The result of a model's prediction: whether a user will click on an ad (1 with probability $p = 0,25$) or not (0 with probability 0,75)
БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ	Binomial Distribution is the distribution of the number of successes in a sequence of an independent Bernoulli trials with the same probability of success p . Binomial distribution is used for A/B testing and classifier accuracy estimation	The probability that exactly 15 out of 100 website users will make a purchase with a 12 % conversion rate is described by the binomial distribution
НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ	Normal Distribution is a continuous symmetric distribution about the mean value, characterized by parameters μ (mean) and σ^2 (variance). Normal distribution is fundamental for statistical inference and many machine learning algorithms	The height of mobile app users follows a normal distribution with a mean of 170 cm and a standard deviation of 10 cm
РАСПРЕДЕЛЕНИЕ ПУАССОНА	Poisson Distribution is a distribution describing the number of events occurring over a fixed time interval, assuming events occur independently with a constant average intensity. Poisson distribution models rare events	The number of server failures per hour (averaging 0,5 failures) is described by the Poisson distribution – this is used for resource planning and SLA

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ	Uniform Distribution is a distribution where all outcomes are equally likely. Uniform distribution is used for initializing neural network weights and generating random numbers	When initializing neural network weights, values are selected from a uniform distribution on the interval $[-0,1; 0,1]$
ЗАКОН БОЛЬШИХ ЧИСЕЛ	Law of Large Numbers is a theorem stating that the sample mean converges to the expected value as the sample size increases. The Law of Large Numbers justifies the need for big data to train AI models	Model prediction accuracy stabilizes as the test sample increases: on 10 examples accuracy may be 70 %; on 1000 examples – 84 %; on 10 000 examples – 85,2 % (approaching the true value)
ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА	Central Limit Theorem is a theorem stating that the sum of a large number of independent random variables has an approximately normal distribution, regardless of the distributions of individual summands. CLT justifies the use of normal distribution in statistical tests and confidence intervals	The average app rating in the App Store (based on thousands of reviews) has a normal distribution, even if individual ratings are unevenly distributed – this allows building confidence intervals for the average rating

РАЗДЕЛ 2

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ И ОПИСАТЕЛЬНАЯ СТАТИСТИКА

2.1. Принципы описания данных

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ	An approach to analyzing datasets to summarize their main characteristics, often using visual methods. EDA helps understand data structure, identify anomalies, check assumptions, and determine important variables before building machine learning models	In a data scientist's practice, exploratory analysis helps identify outliers and anomalies before training models: for example, when analyzing transaction data, one can detect erroneous purchase amounts (negative numbers or unrealistically large values) that need to be corrected or removed
ОПИСАТЕЛЬНАЯ СТАТИСТИКА	Methods of summarizing and describing the main characteristics of a dataset using numerical indicators and graphs. Descriptive statistics includes calculation of measures of central tendency (mean, median, mode), measures of spread (variance, standard deviation, range), and visualization of distributions	When analyzing an online store's customer database, descriptive statistics show: average check is 2500 rubles, median check is 1800 rubles, standard deviation is 1200 rubles, indicating significant variation in purchase amounts
МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ	Statistics describing the center of data distribution or typical value in a dataset. Main measures: arithmetic mean, median, and mode. Choice of measure depends on data type and presence of outliers	When analyzing IT company employee salaries, the average salary may be 150,000 rubles, but due to high top management salaries, the median is 95,000 rubles, better reflecting the typical income level of most employees

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ	The sum of all values divided by their count. The arithmetic mean is sensitive to outliers and extreme values, which can distort the representation of typical value in data with asymmetric distribution	Mobile app response time per day: if values of 120, 130, 125, 140 ms and one failure – 2500 ms are recorded, the mean will be 603ms, which does not reflect typical system performance
МЕДИАНА	The middle value of an ordered sample: if the number of observations is odd, the median is the central value; if even – the average of two central values. The median is robust to outliers and better characterizes typical value in asymmetric distributions	When analyzing food courier delivery times per shift: out of 100 deliveries, 95 took 20–40 minutes, and 5 deliveries took over 90 minutes due to complex orders. Median time of 28 minutes better reflects typical delivery speed than the mean of 35 minutes
МОДА	The most frequently occurring value in a dataset. Mode is especially useful for categorical data and discrete variables. In multimodal distributions, there can be several modes	In web service error logs for a week, the most common error (mode) is “Timeout database connection” (occurs 247 times), indicating a database connection pool problem and helping prioritize fixes
МЕРЫ РАЗБРОСА	Statistics describing data variability, i.e., how values are scattered relative to the center. Main measures: range, variance, standard deviation, interquartile range (IQR)	Two development teams show the same average task completion speed – 8 story points per sprint. But the first team has standard deviation 1,5 points (stable work), and the second – 4 points (unpredictable performance), which is critical for release planning
РАЗМАХ	The difference between maximum and minimum values in a dataset. Range gives a simple estimate of variability but is extremely sensitive to outliers and does not account for value distribution within the range	Website page load time range: minimum time 0,8 seconds, maximum 45 seconds (due to rare failures). Range of 44,2 seconds creates a false impression of instability, while 99 % of requests load in 0,8–2,5 seconds

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
МЕЖКВАРТИЛЬНЫЙ РАЗМАХ	The difference between the third (Q_3 , 75th percentile) and first (Q_1 , 25th percentile) quartiles. IQR shows the spread of the middle 50 % of data and is robust to outliers, making it a preferred measure of spread for asymmetric distributions	When analyzing Python junior developer salaries: $Q_1 = 60,000$ rub.; $Q_3 = 95,000$ rub.; $IQR = 35,000$ rub. This means the middle 50 % of specialists earn salaries in the range of 60–95 thousand, giving a realistic view of the market without influence of extremely high and low values
КВАНТИЛИ	Points dividing an ordered sample into equal parts. A quantile of level p divides data so that $p \cdot 100$ % of observations lie below this value. Special cases: quartiles (division into 4 parts), percentiles (division into 100 parts)	The 95th percentile of API response time is 450 ms: this means 95 % of requests are processed faster than 450 milliseconds. This indicator is used in SLA (Service Level Agreement) instead of average response time, as it better reflects the experience of most users
ПРОЦЕНТИЛИ	Division of ordered data into 100 equal parts. The n th percentile shows the value below which n % of observations fall. Percentiles are widely used for system performance analysis and setting threshold values	The 90th percentile of mobile app load time is 2,3 seconds: 90 % of users see the fully loaded interface faster than this value, while the remaining 10 % face slower loading due to slow internet or old devices
КВАРТИЛИ	Division of ordered data into 4 equal parts. First quartile (Q_1) corresponds to 25th percentile, second quartile (Q_2) – median (50th percentile), third quartile (Q_3) – 75th percentile. Quartiles describe data distribution and are used to identify outliers.	Analysis of student scores on a machine learning exam: $Q_1 = 62$ points; Q_2 (median) = 74 points; $Q_3 = 85$ points. This shows that 25 % of students scored less than 62 points, 50 % – less than 74 points, and 75 % – less than 85 points, helping assess exam difficulty
АСИММЕТРИЯ РАСПРЕДЕЛЕНИЯ	A measure of distribution asymmetry relative to the mean value. Positive	The income distribution of premium subscription users of a streaming service has

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	skewness (right-sided) indicates a long right tail of distribution (most values concentrated on the left), negative – left-sided asymmetry	positive skewness: most pay the base price of 299 rubles per month, but a small group of corporate clients pays tens of thousands of rubles, creating a long right tail
ЭКЦЕСС	A measure of the sharpness of the distribution peak and tail heaviness compared to normal distribution. High kurtosis (positive) indicates heavy tails and a sharp peak (more outliers), low kurtosis (negative) a flat distribution with light tails	The daily stock return distribution of a technology company has high kurtosis: most days show small variability (sharp peak near 0 %), but periodically there are sharp price jumps of 10–15 % (heavy tails), which is critical for risk assessment
ВЫБРОС	An observation that differs significantly from other observations in the sample. Outliers may indicate data errors, rare events, or important anomalies. In machine learning, outliers can distort models and require special handling	In server monitoring systems, a processor temperature of 125 °C was recorded with typical values of 45–70 °C. This outlier may indicate a cooling system failure, sensor error, or extreme load – immediate verification is required to prevent equipment damage
ЯЩИК С «УСАМИ»	A graphical representation of a five-number summary of data: minimum value, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum value. The box covers the interquartile range (IQR), “whiskers” show the data range, and points beyond whiskers are outliers	When comparing the performance of three classification algorithms, boxplots show: algorithm <i>A</i> has median accuracy 87 % with small spread, algorithm <i>B</i> – 89 % but with larger spread, algorithm <i>C</i> – 85 % with several downward outliers. This helps choose a stable algorithm for production
ГИСТОГРАММА	A graphical representation of numerical data distribution: the value range is divided into intervals (bins), and for each interval the frequency (count) of hits	A histogram of user session times on a website shows a bimodal distribution: first peak at 2–3 minutes (quick view and leave), second peak at 15–20 minutes

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	is displayed. Histogram visualizes distribution shape, modality, and asymmetry	(engaged users). This helps segment the audience and optimize content
ДИАГРАММА РАССЕЯНИЯ	A graph for displaying the relationship between two numerical variables: each point represents an observation with coordinates (x, y) . Scatter plot helps identify linear or nonlinear dependencies, clusters, and outliers in two-dimensional data	A scatter plot of apartment price vs. area dependence shows positive correlation: as area increases, price increases. Outliers are also visible – elite apartments with abnormally high price per square meter and clusters by city districts
КОРРЕЛЯЦИЯ	A statistical relationship between two variables showing how change in one variable is related to change in another. Correlation does not imply causation (correlation does not imply causation)	A strong positive correlation was found between the number of YouTube video views and the number of likes ($r = 0,78$). However, this does not mean likes cause views: both indicators depend on content quality and author popularity
КОРРЕЛЯЦИЯ ПИРСОНА	A measure of linear correlation between two variables, taking values from -1 to $+1$. Value $+1$ indicates perfect positive linear dependence; -1 – perfect negative; 0 – no linear connection. Requires normal data distribution	Pearson correlation between time spent studying a <i>data science</i> course and final grade is $r = 0,65$ (moderate positive correlation): students who spent more hours on material tend to receive higher grades
КОРРЕЛЯЦИЯ СПИРМЕНА	A measure of monotonic (not necessarily linear) relationship between variables, based on value ranks rather than values themselves. Spearman correlation is used for non-normally distributed data, ordinal scales, and in the presence of outliers	When analyzing employee satisfaction (ratings from 1 to 5) and their productivity (number of completed tasks), Spearman correlation $\rho = 0,42$ shows moderate positive relationship: more satisfied employees tend to work more productively, even if the dependence is nonlinear

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
КОВАРИАЦИЯ	A measure of joint variability of two random variables showing the direction of their linear relationship. Positive covariance indicates variables change in the same direction, negative – in opposite directions. Unlike correlation, covariance depends on data scale	The covariance between a company's stock price and trading volume is positive: when price rises, trading activity usually increases. However, the absolute covariance value is difficult to interpret without normalization, so correlation coefficient is used to compare relationship strength
ВИЗУАЛИЗАЦИЯ ДАННЫХ	Graphical representation of data and information using visual elements (graphs, diagrams, maps, dashboards). Visualization helps quickly identify patterns, anomalies, trends, and relationships that are difficult to detect when analyzing raw data	An interactive dashboard in Tableau shows key metrics of an online store: sales dynamics, order geography, funnel conversion, top products. Managers assess the situation and make decisions in seconds instead of studying tables with thousands of rows
ПРОПУЩЕННЫЕ ДАННЫЕ	Missing values in a dataset that may occur due to collection errors, users refusing to answer questions, or equipment failures. Missing data leads to biased estimates and reduced model quality if not handled correctly	In a customer income survey, 30 % of respondents did not specify the amount. Simply deleting these rows will bias the sample toward more open customers. Better to use Multiple Imputation or add a missing indicator as a separate feature
ОЧИСТКА ДАННЫХ	The process of detecting and correcting errors in data: handling missing values, removing duplicates, correcting incorrect formats, eliminating outliers, standardizing categories. Data cleaning takes 60–80 % of time in data science projects	When preparing transaction data for an ML model, cleaning includes: removing test transactions with 0 ruble amount, correcting dates in wrong format, merging duplicate categories ("Moscow" and "moscow"), replacing outliers (10 million ruble purchases) with maximum realistic value

2.2. Статистическое оценивание и проверка гипотез

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
СТАТИСТИЧЕСКИЙ ВЫВОД	The process of using sample data to draw conclusions about the population: estimating parameters, testing hypotheses, building confidence intervals. Statistical inference allows generalizing experiment results to the entire population considering uncertainty	Based on a survey of 1500 mobile app users (sample), we conclude about the entire base of 2 million users (population): $68 \pm 3 \%$ are satisfied with the new interface with 95 % confidence
ТОЧЕЧНОЕ ОЦЕНИВАНИЕ	Estimation of a population parameter by a single number based on a sample. Examples: sample mean (estimate of expected value), sample variance (estimate of population variance), sample proportion (estimate of event probability)	The average courier delivery time for a week was 32 minutes – this is a point estimate of the average delivery time for the entire population of orders. However, point estimate does not show uncertainty
ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ	Estimation of a parameter as an interval of values (confidence interval) that contains the true parameter value with a given probability. Interval estimation is more informative than point estimation as it shows estimate uncertainty	Confidence interval for landing page conversion: $12,5 \pm 1,8 \%$ (at 95 % confidence level). This means the true conversion of the population lies between 10,7 and 14,3 %, not exactly 12,5 %
ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ	An interval of values that contains the true value of the estimated parameter with a given confidence probability. Interval width depends on sample size, data variability, and chosen confidence level	95 % confidence interval for average Python developer salary in Moscow: $145,000 \pm 12,000$ rubles. Interpretation: if we repeatedly take samples and build intervals, 95 % will contain the true mean salary

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ	The probability that the confidence interval contains the true parameter value. Common confidence levels: 90, 95, 99 %. Higher confidence level gives wider interval (more uncertainty) but greater reliability	In <i>A/B</i> testing a new feature, we use 95 % confidence level: if the confidence interval of conversion difference does not include zero, we conclude statistical significance of the effect
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ	A procedure for testing statements about population parameters based on sample data. Includes formulating null and alternative hypotheses, choosing significance level, calculating test statistic, and deciding to reject or not reject null hypothesis	Testing hypothesis about new recommendation algorithm efficiency: null hypothesis H_0 – average check unchanged, alternative H_1 – average check increased. After collecting data and calculating p -value = 0,023, we conclude statistical significance
НУЛЕВАЯ ГИПОТЕЗА, H_0	The hypothesis being tested and that researchers try to reject. Null hypothesis usually states no effect, no differences, or no relationship between variables (“status quo”). In ML, H_0 often states new model is not better than baseline	When comparing two classification models, null hypothesis: new model accuracy equals current production model accuracy (difference = 0). Alternative: new model is more accurate. If p -value < 0,05, we reject H_0
АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА, H_1	The hypothesis contradicting the null. Alternative hypothesis states presence of effect, differences, or relationship. Can be one-sided (parameter greater or less) or two-sided (parameter not equal)	In testing a new website button, alternative hypothesis: new version conversion is higher than current (one-sided test). If data shows statistically significant increase with p -value = 0,01, we accept H_1
УРОВЕНЬ ЗНАЧИМОСТИ, α	Probability of rejecting the null hypothesis when it is true (Type I error). Significance level is the threshold for p -value	When launching <i>A/B</i> test, we set α = 0,05. If p -value = 0,03 (< 0,05), we reject null hypothesis about no difference and

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	below which result is considered statistically significant. Standard $\alpha = 0,05$ (5 %)	conclude effect significance. Risk: 5 % chance we falsely conclude effect exists
P-ЗНАЧЕНИЕ	Probability of obtaining observed or more extreme results assuming the null hypothesis is true. Small p -value (less than significance level α) indicates data incompatibility with H_0 and is grounds for rejection	After comparing two algorithms, we got p -value = 0,008. This means: if algorithms were equally effective (H_0 true), probability of observing such accuracy difference would be only 0,8 %. We conclude significant difference
ОШИБКА ПЕРВОГО РОДА	Rejection of a true null hypothesis (<i>false positive</i>). Type I error probability equals significance level α . In business, Type I error is implementing a change that doesn't actually work, leading to wasted costs	Launching a new app feature showed statistically significant engagement increase ($p < 0,05$), we deployed it. A month later metrics returned to baseline: initial effect was false positive (Type I error)
ОШИБКА ВТОРОГО РОДА	Acceptance (failure to reject) of a false null hypothesis (<i>false negative</i>). Type II error means missing a real effect. Type II error probability is denoted β . Test Power = $1 - \beta$	New fraud detection algorithm is really 15 % more accurate, but due to small sample (only 200 transactions) test didn't find significance ($p = 0,12$). We rejected effective algorithm – this is Type II error
МОЩНОСТЬ ТЕСТА	Probability of correctly rejecting a false null hypothesis ($1 - \beta$). Power shows test's ability to detect an effect if it truly exists. For reliable results, power should be at least 0,8 (80 %)	Before launching A/B test, we calculate required sample size for power 0,8 at expected effect 5 % and $\alpha = 0,05$. We get: need minimum 2000 users per group. Smaller sample risks missing real effect
ТЕСТОВАЯ СТАТИСТИКА	A statistic calculated from sample to test hypothesis (t -statistic, z -statistic, χ^2 , F -statistic).	When comparing means of two groups, we calculated t -statistic = 2,47. From Student's t -distribution

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	Test statistic is compared with critical value or used to calculate p-value. Distribution is known when H_0 is true	table for $\alpha = 0,05$ and $df = 58$, we find critical value $t_{crit} = 2,00$. Since $ 2,47 > 2,00$, we reject H_0
КРИТИЧЕСКАЯ ОБЛАСТЬ	The set of test statistic values for which the null hypothesis is rejected. Critical region is determined by significance level α and alternative hypothesis (one-sided or two-sided)	For two-sided z-test at $\alpha = 0,05$, critical region: $z < -1,96$ or $z > +1,96$. Calculated z-statistic = 2,31 falls in critical region ($2,31 > 1,96$), so we reject H_0 at 5 % significance level
T-ТЕСТ	Statistical test for comparing means using Student's t -distribution. Applied when population variance is unknown and estimated from sample (especially important for small samples $n < 30$)	We compare API response time before and after optimization: sample of 25 measurements showed t -statistic = 3,12 at $df = 24$. p -value = 0,004 < 0,05, we conclude statistically significant performance improvement
Z-ТЕСТ	Test based on normal distribution (z-distribution). Applied for large samples ($n \geq 30$) or when population variance is known. For proportions and means at large n , z-test and t -test give similar results	We test hypothesis about landing page conversion: out of 5000 visitors, 625 completed target action (12,5 %). With known variance and large n , we use z-test. z-statistic = 1,89; p -value = 0,059 – difference with target 11 % is not significant at 5 %
ХИ-КВАДРАТ-ТЕСТ	Test for checking independence of two categorical variables or fit to theoretical distribution. χ^2 test compares observed frequencies with expected frequencies under null hypothesis	We check dependence between user gender and tariff choice (basic/premium). Contingency table 2×2 gives $\chi^2 = 8,73$, $df = 1$, p -value = 0,003. We reject independence hypothesis: men and women differ significantly in tariff choice
ДИСПЕРСИОННЫЙ АНАЛИЗ	Method for comparing means of three or more groups simultaneously. ANOVA tests null	We compare efficiency of three marketing channels (email, social, context) by average customer

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	hypothesis about equality of all group means. If ANOVA shows significance, post-hoc tests (Tukey, Bonferroni) are conducted for pairwise comparisons	acquisition cost. One-way ANOVA: $F(2, 87) = 5,34; p = 0,006$. Channels differ significantly. Post-hoc Tukey shows: context is more expensive than email
F-ТЕСТ	Test using F -distribution for comparing variances or in ANOVA for comparing multiple means. F -statistic is ratio of two variances. When comparing variances, F -close to 1 indicates equality	In ANOVA for comparing four ML models: F -statistic = 4,67; $df_{\text{between}} = 3$; $df_{\text{within}} = 96; p = 0,005$. Models differ significantly in accuracy. Gradient boosting shows highest mean accuracy (92,3 %)
ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	Number of independent values in statistic calculation, equal to sample size minus number of estimated parameters. Degrees of freedom affect shape of t -distribution, χ^2 -distribution, and F -distribution	For t -test of two independent samples size $n_1 = 30$ and $n_2 = 35$, degrees of freedom: $df = n_1 + n_2 - 2 = 63$. From t -distribution table for $df = 63$ and $\alpha = 0,05$ (two-sided), we find critical value $t_{\text{crit}} = 1,998$
РАЗМЕР ЭФФЕКТА	Quantitative measure of the strength of the studied phenomenon, independent of sample size. Popular measures: Cohen's d (for mean difference), η^2 (eta-squared for ANOVA), odds ratio (for binary outcomes)	We compared two models: new model is significantly more accurate ($p = 0,001$), but Cohen's $d = 0,12$ (very small effect). Accuracy difference 0,3 % is significant due to large sample, but practically insignificant for business
РАЗМЕР ВЫБОРКИ	Number of observations in a sample. Larger sample size provides more accurate estimates, narrow confidence intervals, and high test power. However, after certain limit accuracy gain becomes insignificant	To detect 2 % conversion increase (from 10 % to 12 %) at power 0,8 and $\alpha = 0,05$, needed sample size is 3853 observations per group. Smaller sample (e.g., 500) won't detect effect even if real (low power)

РАЗДЕЛ 3

СТАТИСТИЧЕСКИЕ МОДЕЛИ ДЛЯ ОБУЧЕНИЯ С УЧИТЕЛЕМ

3.1. Регрессия со статистической точки зрения

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ОБУЧЕНИЕ С УЧИТЕЛЕМ	Type of machine learning where algorithm learns on labeled data, i.e., data with known correct answers (target variables). Model learns to find relationship between input features and output value to make predictions for new data	Bank credit scoring system learns on historical borrower data: for each client features are known (age, income, credit history) and target variable (repaid credit or not). After training, model predicts default probability for new applicants
РЕГРЕССИЯ	Method of modeling relationship between dependent (target) variable and one or more independent variables (features). Regression analysis is used for predicting continuous numerical values, not categorical labels	“Yandex.Real” Estate service uses regression to predict apartment prices: based on features (area, floor, district, year built) model predicts market price in rubles – a continuous numerical value
ЗАВИСИМАЯ ПЕРЕМЕННАЯ	The variable that the model predicts or explains; also called target variable or response. Dependent variable “depends” on values of independent variables (features)	In forecasting demand for goods in Wildberries store, dependent variable is the number of sales of specific product per day – exactly this value we want to predict based on features
НЕЗАВИСИМАЯ ПЕРЕМЕННАЯ	Variable used for predicting dependent variable; also called feature, predictor, or explanatory variable.	When predicting “Yandex.Food” delivery time, independent variables are: distance from restaurant to client, time

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	Independent variables are considered model input data	of day, day of week, weather, number of orders with courier – all these features affect target variable
ПРИЗНАК	A single measurable characteristic of an observed phenomenon; a single column in a dataset. Features can be numerical (age, income), categorical (gender, city), or binary (discount available: yes/no)	In email spam recognition system, features include: number of words in email, presence of certain keywords (“win”, “free”), number of links, sender domain, sending time, attachments – each feature helps model distinguish spam
ЦЕЛЕВАЯ ПЕРЕМЕННАЯ	The variable that the model tries to predict; synonym of dependent variable. In classification tasks target variable takes categorical values (classes), in regression tasks – continuous numerical values	In Sberbank fraud detection task, target variable takes two values: 1 (fraud) or 0 (legitimate operation). Model analyzes transaction features and predicts probability of belonging to fraud class
ЛИНЕЙНАЯ РЕГРЕССИЯ	Regression analysis modeling linear relationship between dependent variable and one or more independent variables. Linear regression assumes change in dependent variable is proportional to change in predictors	Model predicts car fuel consumption based on distance traveled: if car consumes 8 liters per 100 km, linear dependence is $y = 0,08x$, where x is distance in km, y is fuel consumption in liters
ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ	Regression with one independent variable (one feature). Simple linear regression is described by equation $y = \beta_0 + \beta_1 x + \varepsilon$, where β_0 is intercept, β_1 is slope coefficient, ε is random error	Analysis of dependence between air temperature and ice cream sales at a kiosk: when temperature rises by 1 °C, sales increase by average 15 portions. Regression equation: $\text{sales} = 50 + 15 \times \text{temperature}$, where 50 is baseline sales at 0 °C

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ	Regression with several independent variables (several features). Multiple regression is described by equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$, allowing to account for influence of multiple factors simultaneously	Courier delivery cost forecasting considers several features: distance (km), order weight (kg), time of day (hour), day of week, weather conditions. Equation: $\text{cost} = 100 + 15 \times \text{distance} + 5 \times \text{weight} + 20 \times \text{night tariff}$
КОЭФФИЦИЕНТЫ РЕГРЕССИИ	Model parameters ($\beta_1, \beta_2, \dots, \beta_n$) determining contribution of each independent variable to dependent variable prediction. Coefficient shows by how many units target variable changes when corresponding predictor changes by one unit	In developer salary prediction model, regression coefficients show: work experience (years) has coefficient +15 000 (each year adds 15k rubles); English knowledge gives +25 000; working in Moscow +30 000, remote work –10 000
СВОБОДНЫЙ ЧЛЕН	Value of dependent variable at zero values of all independent variables; point where regression line intersects Y-axis. Intercept represents baseline value when all predictors are zero	In taxi cost model intercept = 150 rubles means base car submission cost charged regardless of trip distance. Even if distance is 0 km (passenger didn't go), minimum payment is 150 rubles
ОСТАТОК	Difference between observed (actual) value of dependent variable and value predicted by model: $e_i = y_i - \hat{y}_i$. Residuals show how well model describes each specific observation; residual analysis helps check model assumptions	Model predicted delivery time 35 minutes, but actual delivery took 42 minutes. Residual equals $42 - 35 = +7$ minutes (model underestimated time). For another delivery residual was –3 minutes (model overestimated)
СУММА КВАДРАТОВ ОСТАТКОВ	Sum of squared differences between observed and predicted values: $\text{SSE} = \sum (y_i - \hat{y}_i)^2$.	When comparing two housing price prediction models: first model has $\text{SSE} = 2,500,000$,

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	SSE is a measure of total model error; Ordinary Least Squares (OLS) method finds regression coefficients minimizing this sum	second SSE = 1,800,000. Second model is better as sum of squared errors is smaller – predictions are closer to actual prices
КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ	Proportion of dependent variable variance explained by the model; measure of model fit quality. R^2 takes values from 0 to 1: $R^2 = 0,85$ means 85 % of target variable variation is explained by model features	Movie rating prediction model on Kinopoisk has $R^2 = 0,72$: 72 % of differences in user ratings are explained by features (genre, budget, director rating, actors), and 28 % are due to subjective factors model doesn't account for
СКОРРЕКТИРОВАННЫЙ R^2	Modification of R^2 adjusted for number of predictors in model. Unlike regular R^2 which always grows when adding new features, Adjusted R^2 penalizes adding useless predictors that don't substantially improve model	When adding “programmer's favorite color” feature to salary prediction model, regular R^2 grew from 0,78 to 0,781 (insignificantly), but Adjusted R^2 decreased from 0,76 to 0,758 – signal that new feature carries no useful information
СРЕДНЕ-КВАДРАТИЧНАЯ ОШИБКА	Mean value of squared errors: $MSE = \left(\frac{1}{n}\right)\sum(y_i - \hat{y}_i)^2.$ MSE is widely used as loss function in regression; quadratic nature amplifies penalty for large errors	Comparison of two demand forecasting algorithms: first has MSE = 144, second MSE = 100. Second algorithm is more accurate: mean squared error is smaller. MSE is measured in squares of target variable units
СРЕДНЕ-КВАДРАТИЧНОЕ ОТКЛОНЕНИЕ	Square root of MSE: $RMSE = \sqrt{MSE}$. RMSE is measured in same units as target variable, which simplifies interpretation. RMSE shows	Delivery time prediction model has $RMSE = 8$ minutes: this means typical deviation of prediction from actual time is about 8 minutes. If

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	“typical” magnitude of prediction error	average delivery time is 35 minutes, most predictions fall within ± 8 minutes range
СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА	Mean value of absolute errors: $MAE = \left(\frac{1}{n}\right) \sum y_i - \hat{y}_i .$ MAE is less sensitive to outliers than MSE as it doesn't square errors; MAE is also measured in original units of target variable	Two apartment price prediction models: first has MAE = 350,000 rubles; MSE = 200 billion; second MAE = 380,000 rubles; MSE = 180 billion. First model averages less error (MAE), but second rarely makes catastrophically large errors (MSE)
ПЕРЕОБУЧЕНИЕ	Situation when model fits training data too well, including their noise and random fluctuations, and loses ability to generalize to new data. Overfitted model shows excellent results on training data but poor on test data	Cat and dog recognition model memorized all 1000 training photos by heart (100 % accuracy), but on new photos accuracy dropped to 60 %. Model memorized specific images rather than learning general animal features – this is overfitting
НЕДООБУЧЕНИЕ	Situation when model is too simple for data and cannot capture main patterns. Underfitted model shows poor results on both training and test data; model has high bias	Attempt to describe complex seasonal ice cream sales (summer peaks, winter dips, weekend growth) with a straight line gives 45 % accuracy – model is too simple and doesn't account for seasonality, day of week, weather
КОМПРОМИСС СМЕЩЕНИЯ И ДИСПЕРСИИ	Fundamental compromise in machine learning: reducing bias usually increases variance and vice versa. Ideal model balances between underfitting (high bias)	Simple linear regression for housing prices has high bias (doesn't capture nonlinearities) but low variance (stable on different samples). Complex neural network

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	and overfitting (high variance)	has low bias (accurately describes training data) but high variance. Optimal is medium complexity model
ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ	Regression with polynomial features (powers of original variables), allowing to model nonlinear dependencies. Polynomial regression remains linear in parameters but includes features x^2 , x^3 , etc.	Dependence between car speed and fuel consumption is nonlinear: at low speeds consumption grows slowly, at medium – optimal, at high – sharply increases. Second-degree polynomial regression (with features speed and speed ²) accurately describes this U-shaped dependence
РЕГУЛЯРИЗАЦИЯ	Method of preventing overfitting by adding penalty term to loss function limiting model coefficients magnitude. Regularization reduces model complexity, improving generalization ability	When training customer churn prediction model, L2 regularization (Ridge) was added. Model coefficients became smaller in magnitude (from 0,01 to 0,5 instead of –5 to +10 without regularization), test accuracy grew from 78 to 84 %
ГРЕБНЕВАЯ РЕГРЕССИЯ	L2 regularization method adding penalty for sum of squared coefficients to loss function: $\lambda \cdot \sum_{j=1}^p \beta_j^2$ Ridge regression shrinks coefficients but doesn't zero them completely; all features remain in model	In movie rating prediction task with 100 features (genres, actors, director, budget, etc.) Ridge regression reduced all coefficients proportionally but kept all features. Validation accuracy 82 vs 76 % without regularization – overfitting eliminated
ЛАССО-РЕГРЕССИЯ	L1 regularization method adding penalty for sum of absolute coefficient values:	When predicting car insurance cost from 50 features, Lasso zeroed 23 coefficients (kept only significant:

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	$\lambda \cdot \sum_{j=1}^p \beta_j^2$ Lasso can zero some coefficients, performing automatic feature selection	driver age, experience, car brand, engine power, region, accident history). Model became simpler and more interpretable
ЭЛАСТИЧНАЯ СЕТЬ	Combination of $L1$ and $L2$ regularization combining advantages of Ridge and Lasso. Elastic Net is especially useful with high feature correlation (multicollinearity) and when number of features exceeds observations	In genomic analysis for disease prediction from 20,000 genes (features) with only 500 patients, Elastic Net selected 150 significant genes ($L1$ component) and stabilized coefficients ($L2$ component). Classification accuracy 89 vs 76 % for pure Lasso
МУЛЬТИ-КОЛЛИНЕАРНОСТЬ	High correlation between independent variables (features) in regression model. Multicollinearity makes interpreting individual coefficients difficult (impossible to separate influence of correlated features) and increases estimate variance	In apartment price prediction model, features “area in square meters” and “area in square feet” have correlation 0,99 (complete multicollinearity). Coefficients become unstable: when adding/removing other features, coefficients change sign. Need to remove one feature
ГОМО-СКЕДАСТИЧНОСТЬ	Constancy of error (residual) variance across all levels of independent variables; one of key linear regression assumptions. Under homoscedasticity, residual spread is same for all predicted values	In salary prediction model residual plot shows: point spread (residuals) is approximately same for all levels of predicted salaries (from 40 to 300k rub.) – error variance is constant, homoscedasticity assumption holds
ГЕТЕРО-СКЕДАСТИЧНОСТЬ	Non-constancy of error variance: residual spread changes depending on predictor values or predicted values.	Household expense prediction model shows: for poor families (income 30k rub) residuals are small (± 2k rub), for rich

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	Heteroscedasticity leads to inefficient estimates and biased standard errors, distorting hypothesis testing	(income 300k) spread is huge ($\pm 50k$ rub). This is heteroscedasticity: error variance grows with income

3.2. Вероятностная классификация и оценка моделей

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
КЛАССИФИКАЦИЯ	Task of predicting categorical class label for new observation based on training data. Unlike regression (number prediction), classification assigns object to one of predefined classes (categories)	Bank security system classifies each transaction: “fraud” or “legitimate”. Email service classifies messages: “spam” or “not spam”. Medical system diagnoses: “healthy” / “sick” / “needs additional tests”
БИНАРНАЯ КЛАССИФИКАЦИЯ	Classification with two classes (positive and negative). Binary classification is most common task type: spam/not spam, sick/healthy, default/solvent, click/no click	Model determines whether user will click on advertising banner (class 1) or ignore (class 0). Based on features (click history, time on site, device, traffic source), algorithm predicts click probability
МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ	Classification with three or more classes. Multiclass classification requires special algorithms or “one-vs-rest” or “one-vs-one” strategies	MNIST handwritten digit recognition: digit image (28×28 pixels) is classified as one of 10 classes (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). News classification by topics: sports, politics, economy, culture, technology (5 classes)
ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ	Statistical model for binary classification predicting probability of belonging to positive	Model predicts credit application approval probability: based on features (income,

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	class. Logistic regression uses sigmoid function to transform linear combination of features into probability from 0 to 1	credit history, age, loan amount) logistic regression outputs probability 0,73 (73 %). If threshold is 0,5, application is approved
СИГМОИДА	<i>S</i> -shaped function mapping any input values to range from 0 to 1: $\sigma(z) = 1/(1 + e^{-z})$. Sigmoid transforms linear combination of features into probability, making it ideal for logistic regression and neural networks	Logistic regression calculated $z = 2,5$ (linear combination of features). Sigmoid $\sigma(2,5) = 0,92$ – model predicts 92 % probability of positive class. If $z = -1,5$, $\sigma(-1,5) = 0,18$ – only 18 % probability
ГРАНИЦА ПРИНЯТИЯ РЕШЕНИЯ	Boundary in feature space separating classes; surface on one side of which model predicts one class, on the other – another class. In logistic regression boundary is linear (straight line or hyperplane)	In iris classification task by two features (petal length and width), logistic regression built linear boundary: if $2 \times \text{length} + 3 \times \text{width} > 15$, we predict Iris Virginica, otherwise Iris Setosa
ВЕРОЯТНОСТНЫЙ ПОРОГ	Threshold probability value for assigning object to positive class. By default threshold is usually set at 0,5, but can be shifted depending on cost of different error types	Breast cancer diagnosis model outputs probability 0,47. At standard threshold 0,5 result is “healthy”, but since false negative (missing cancer) is dangerous, threshold is lowered to 0,3. Now probability 0,47 is classified as “needs examination”
МАТРИЦА ОШИБОК	$K \times K$ table (where K is number of classes) for classification quality evaluation, showing number of true positives, true negatives,	Pneumonia diagnosis model by X-ray on 100 patients showed: 45 true positives (TP), 40 true negatives (TN), 5 false positives (FP),

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	false positives, and false negatives	10 false negatives (FN). Confusion matrix helps calculate accuracy, precision, recall
ИСТИННО ПОЛОЖИТЕЛЬНЫЙ	Correctly predicted positive class: model predicted positive class, and this matches reality	Fraud detection system correctly identified 247 fraudulent transactions out of 300 real ones (TP = 247). These are good results: most real fraud cases detected
ИСТИННО ОТРИЦАТЕЛЬНЫЙ	Correctly predicted negative class: model predicted negative class, and this matches reality	Spam filter correctly passed 1 850 legitimate emails to “Inbox” folder (TN = 1 850). Important emails are not lost, users are satisfied with filter work
ЛОЖНОПОЛОЖИТЕЛЬНЫЙ	Incorrectly predicted positive class (Type I error): model predicted positive class, but object actually belongs to negative class	Pregnancy test showed positive result, but woman is not pregnant (FP). In spam filtering, FP is when important work email is mistakenly marked as spam and went to “Spam” folder
ЛОЖНООТРИЦАТЕЛЬНЫЙ	Incorrectly predicted negative class (Type II error): model predicted negative class, but object actually belongs to positive class	COVID-19 test showed negative result for infected person (FN = false negative). Person considers themselves healthy, doesn’t isolate and infects others. In medicine, FN is often more dangerous than FP
ТОЧНОСТЬ	Proportion of correct predictions among all predictions:	Rare disease diagnosis model (1 % sick) simply predicts “healthy” for everyone and gets Accuracy = 99 %. But this is useless model: it finds no sick people! For

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	Accuracy = $= \frac{TP + TN}{TP + TN + FP + FN} \cdot$ Accuracy is simple to interpret but can be misleading with imbalanced classes	imbalanced data need Precision, Recall, F1-Score
ПРЕЦИЗИОННОСТЬ	Proportion of true positives among all predicted positives: Precision = $TP / (TP + FP)$. Precision answers: "Among objects model called positive, what proportion is actually positive?"	Spam filter marked 120 emails as spam, of which 108 are really spam (TP = 108), and 12 are important emails (FP = 12). Precision = $108 / 120 = 0,90$ (90 %). High Precision means when filter says "spam", it's usually right
ПОЛНОТА	Proportion of true positives among all actual positives: Recall = $TP / (TP + FN)$. Recall answers: "What proportion of real positive objects did model find?"	In database of 300 fraudulent transactions, model detected 247 (TP = 247), missed 53 (FN = 53). Recall = $247 / 300 = 0,82$ (82 %). Model found 82 % of real fraud. High Recall is critical when missing positive case is dangerous
ЧУВСТВИТЕЛЬНОСТЬ	Synonym of Recall; proportion of true positives among all actual positives. Sensitivity measures model's ability to detect positive cases	HIV test has Sensitivity = $0,995$ (99,5 %): out of 1000 infected people, test will correctly detect 995, and only 5 will get false negative result. High sensitivity is critical for screening tests
СПЕЦИФИЧНОСТЬ	Proportion of true negatives among all actual negatives: $\text{Specificity} = \frac{TN}{TN + FP} \cdot$	Allergy test has Specificity = $0,92$ (92 %): out of 100 healthy people, test will correctly identify 92 as not

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	Specificity measures model's ability to correctly identify negative cases	having allergy, and 8 will get false positive result. High specificity is important when false alarm leads to expensive examinations
F1-МЕРА	Harmonic mean of Precision and Recall: $F1 = \frac{2TP}{2TP + FP + FN}$ F1-Score is useful when balance between Precision and Recall is needed, especially with imbalanced classes	Model A: Precision = 0,95; Recall = 0,60; F1 = 0,74. Model B: Precision = 0,80; Recall = 0,85; F1 = 0,82. Although Model A has higher Precision, Model B is better by F1 – it provides better balance
РОС-КРИВАЯ	Graph showing dependence of True Positive Rate (Recall) on False Positive Rate at various threshold settings. ROC curve visualizes tradeoff between detecting positive cases and false alarms	At threshold 0,9: TPR = 0,50; FPR = 0,05. At threshold 0,7: TPR = 0,75; FPR = 0,15. At threshold 0,5: TPR = 0,90; FPR = 0,30. ROC curve connects these points; closer curve to top-left corner, better model
ПЛОЩАДЬ ПОД РОС-КРИВОЙ (AUC)	Area Under ROC Curve; integral classifier quality metric. AUC takes values from 0,5 to 1,0: AUC = 0,5 means random guessing, AUC = 1,0 means perfect classification	Model A has AUC = 0,92 – excellent quality: in 92 % of cases model correctly ranks positive example higher than negative. Model B: AUC = 0,68 – weak model, only slightly better than random guessing
ДОЛЯ ИСТИННО ПОЛОЖИТЕЛЬНЫХ	Proportion of correctly predicted positive cases among all actual positives:	Out of 200 patients with “pneumonia” diagnosis, model correctly identified 176

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	$TPR = TP / (TP + FN)$. TPR coincides with Recall and Sensitivity	(TP = 176), missed 24 (FN = 24). $TPR = 176 / 200 = 0,88$ (88 %). Model detected 88 % of real pneumonia cases
ДОЛЯ ЛОЖНОПОЛОЖИТЕЛЬНЫХ	Proportion of incorrectly predicted positive cases among all actual negatives: $FPR = FP / (FP + TN)$. FPR shows how often model “falsely alarms”	Out of 1000 healthy people, test gave 80 false positive results (FP = 80), 920 correct negatives (TN = 920). $FPR = 80 / 1000 = 0,08$ (8 %). Test falsely diagnoses disease in 8 % of healthy people
КРОСС-ВАЛИДАЦИЯ	Method of evaluating model generalization ability by repeatedly splitting data into training and validation sets. Cross-validation gives more reliable quality estimate than single train/test split	Instead of single 80/20 split, data is divided into 5 parts (folds). Model is trained 5 times: each time one part is test, other 4 are training. We get 5 accuracy estimates: 0,82; 0,79; 0,84; 0,81; 0,83. Mean accuracy $0,82 \pm 0,02$
К-КРАТНАЯ КРОСС-ВАЛИДАЦИЯ	Cross-validation with data split into k equal parts (folds). 10-fold cross-validation ($k = 10$) is most common, providing good balance between computational cost and evaluation reliability	Dataset of 5000 records is divided into 10 folds of 500 records each. Model is trained 10 times: first time fold 1 is test, folds 2–10 are training; second time fold 2 is test, etc. Final metric is average of 10 runs
СКОЛЬЗЯЩИЙ КОНТРОЛЬ	Cross-validation where each object sequentially becomes test, and all others are training (special case of k -fold where $k = n$, number of objects). LOO uses maximum data for training but is computationally expensive	Dataset of 100 patients: model is trained 100 times. First time: patient 1 is test, patients 2–100 are training. Second time: patient 2 is test, patients 1,3–100 are training. LOO gives least biased estimate but requires 100 trainings

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ОБУЧАЮЩАЯ ВЫБОРКА	Dataset for training (fitting) model; data on which algorithm “learns”, adjusting parameters (weights, coefficients). Usually comprises 60–80% of all data	Out of 10,000 labeled cat and dog images, 8,000 (80 %) are allocated to training set. On these 8,000 images neural network trains for 50 epochs, adjusting weights to minimize loss function
ВАЛИДАЦИОННАЯ ВЫБОРКА	Dataset for hyperparameter tuning and selecting best model configuration; data used for intermediate evaluation during training. Validation set helps prevent overfitting	Out of 10,000 images: 8,000 – training, 1,000 – validation, 1,000 – test. During training after each epoch we check validation accuracy. If training accuracy grows but validation drops – overfitting started, need to stop training (early stopping)
ТЕСТОВАЯ ВЫБОРКА	Dataset for final model evaluation; data not used for either training or hyperparameter tuning. Test set should be used only once to evaluate final model performance	After completing training and hyperparameter tuning on training and validation, final model is evaluated on test set (1,000 images model “hasn’t seen”). Test accuracy 87 % – realistic estimate of how model will work on new data
ГИПЕРПАРАМЕТРЫ	Parameters tuned before model training (unlike model parameters which are learned from data). Examples: learning rate, number of trees in random forest, polynomial degree, regularization coefficient	Gradient boosting has hyperparameters: <code>learning_rate = 0,1</code> ; <code>n_estimators = 100</code> ; <code>max_depth = 5</code> ; <code>min_samples_split = 10</code> . These values are set before training. Grid Search found best values: <code>learning_rate = 0,05</code> ; <code>n_estimators = 200</code> – accuracy grew from 82 to 87 %

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ПОИСК ПО СЕТКЕ	Hyperparameter tuning method by exhaustive search: systematic checking of all combinations of specified hyperparameter values using cross-validation to evaluate each combination	For SVM we specify grid: $C = [0,1; 1; 10]$; $\text{gamma} = [0,001; 0,01; 0,1]$, $\text{kernel} = [\text{'linear'}, \text{'rbf'}]$. Total $3 \cdot 3 \cdot 2 = 18$ combinations. Each combination is evaluated by 5-fold cross-validation (90 model trainings). Best combination: $C = 10$; $\text{gamma} = 0,01$; $\text{kernel} = \text{'rbf'}$ with 91 % accuracy
БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ	Classification based on Bayes theorem, updating prior beliefs with observed evidence. Bayesian methods give probabilistic interpretation of predictions	Prior probability of disease is 1 % ($P(\text{disease}) = 0,01$). Test gave positive result. By Bayes theorem we calculate posterior probability: $P(\text{disease} \text{positive_test}) = 0,32$ (32 %). Despite positive test, disease probability is only 32 % due to false positives
НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР	Classifier with assumption of feature independence given class (conditional independence). Despite “naïve” assumption, Naive Bayes is surprisingly effective for text classification and spam filtering	Spam filter analyzes email: features are presence of words “win” (probability in spam 0,4, in not-spam 0,01); “free” (0,3 vs 0,02); “urgent” (0,25 vs 0,03). Naive Bayes multiplies probabilities and gets: $P(\text{spam} \text{words}) = 0,89$. Email is classified as spam
АПРИОРНАЯ ВЕРОЯТНОСТЬ	Probability of hypothesis (class) before considering data; initial degree of belief in hypothesis based on previous knowledge or class frequency in data	In rare disease diagnosis task prior probability $P(\text{sick}) = 0,01$ (1 % of population is sick); $P(\text{healthy}) = 0,99$. Even with positive test (which has 95 % accuracy) posterior

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
		probability of disease remains low due to small prior probability
АПОСТЕРИОРНАЯ ВЕРОЯТНОСТЬ	Probability of hypothesis (class) after considering data; updated probability calculated by Bayes theorem based on prior probability and data likelihood	Prior spam probability $P(\text{spam}) = 0,30$ (30 % of emails are spam). Email contains word “win” which occurs in 40 % of spam and 1 % of not-spam. By Bayes: $P(\text{spam} \text{“win”}) = 0,95$. After considering word presence, posterior spam probability grew from 30 to 95 %
ПРАВДОПОДОБИЕ	Probability of data given hypothesis; probability of observing available data if hypothesis is true. In Bayes theorem: $P(\text{data} \text{hypothesis})$	Hypothesis H : coin is fair ($P(\text{heads}) = 0,5$). Data D : 8 heads out of 10 tosses. Likelihood $P(D H) = C(10,8) \times 0,5^8 \cdot 0,5^2 = 0,044$ (4,4 %). Low likelihood indicates data is unlikely under fair coin hypothesis – maybe coin is biased

РАЗДЕЛ 4

ПРОДВИНУТЫЕ ТЕМЫ: МЕТОДЫ БЕЗ УЧИТЕЛЯ, МАСШТАБИРУЕМОСТЬ И ИНТЕГРАЦИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

4.1. Статистические подходы к обучению без учителя

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ	Type of machine learning where algorithm works with unlabeled data (without known correct answers). Goal is to discover hidden patterns, structures, or relationships in data independently, without external guidance	Analysis of mobile app user behavior without known categories: algorithm independently discovers there are user groups – “morning active”, “evening passive”, “weekend”, “regular”, though we didn’t know about such segments beforehand
КЛАСТЕРИЗАЦИЯ	Grouping objects by similarity of their characteristics; task of splitting dataset into groups (clusters) so that objects within one cluster are similar to each other, and objects from different clusters differ	Online store customer segmentation by purchasing behavior: cluster 1 – “frequent buyers with low average check”; cluster 2 – “rare buyers with high check”; cluster 3 – “seasonal buyers”; cluster 4 – “churning customers”. Marketing strategy is developed for each cluster
К-СРЕДНИХ КЛАСТЕРИЗАЦИЯ	Iterative clustering method that divides data into K clusters by minimizing within-cluster variance (sum of squared distances from points to	Clustering 10,000 news site articles into 5 thematic groups: K -means algorithm analyzes vector representations of texts

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	their cluster centroids). Algorithm: 1) randomly initialize K centroids; 2) assign points to nearest centroids; 3) recalculate centroids; 4) repeat until convergence	(TF-IDF or word2vec) and automatically groups articles into clusters: politics, sports, economy, culture, technology – without known category labels
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ	Clustering method building hierarchical tree-like structure (dendrogram) of nested clusters. Can be agglomerative (bottom-up: each object is separate cluster, then clusters merge) or divisive (top-down: all objects in one cluster, then split)	Hierarchical clustering of 100 animal species by genetic features: dendrogram shows evolutionary relationship – at top level mammals, birds, reptiles separate; within mammals – predators, herbivores, omnivores; within predators – felines, canines, etc.
ДЕНДРОГРАММА	Tree diagram visualizing hierarchical clustering results; shows sequence of cluster merging or splitting and distances (similarity) between them. Dendrogram branch height reflects degree of difference between clusters	Dendrogram of 50 companies clustering by financial indicators shows: at height 10 units all companies in one cluster; at height 5 – split into 3 large clusters (technology, industrial, financial); at height 2 – within technology, software and hardware companies separate
DBSCAN (Ди Би Скан)	Density-based clustering algorithm that finds clusters of arbitrary shape and identifies outliers (noise). DBSCAN groups together points close to each other (within radius ϵ) and marks isolated points as noise. Parameters: ϵ (neighborhood radius) and minPts (minimum points)	Geolocation anomaly detection: DBSCAN analyzes client transaction coordinates. Main mass of points forms dense cluster (home-work-stores), but several transactions in another city/country are isolated points (noise) – these are anomalies, possible fraud

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ЦЕНТРОИД	Cluster center; point representing mean position of all points belonging to cluster. In <i>K</i> -means centroid is calculated as arithmetic mean of coordinates of all cluster points for each feature	After clustering customers into 4 groups, cluster centroids show typical characteristics: centroid of cluster 1 – age 25, income 40k, purchases 2 times/month; centroid of cluster 2 – age 45, income 120k, purchases 8 times/month
СИЛУЭТНЫЙ КОЭФФИЦИЕНТ	Measure of clustering quality showing how well objects are separated into clusters. For each point calculated: <i>a</i> – mean distance to points in same cluster, <i>b</i> – mean distance to points in nearest neighboring cluster. $\text{Silhouette} = \frac{b - a}{\max(a, b)}$ Values from –1 to +1	Comparison of clustering with <i>K</i> = 3 and <i>K</i> = 5: at <i>K</i> = 3 mean Silhouette Score = 0,65 (good separation), at <i>K</i> = 5 Score = 0,42 (clusters overlap). Conclusion: 3 clusters are more optimal than 5
МЕТОД ЛОКТЯ	Heuristic method for determining optimal number of clusters <i>K</i> in <i>K</i> -means. Graph of within-cluster variance (<i>inertia</i>) dependence on number of clusters is built. Optimal <i>K</i> is at “elbow” point where adding new clusters stops substantially decreasing variance	Graph of inertia from <i>K</i> = 1 to <i>K</i> = 10 shows sharp decrease to <i>K</i> = 4 (inertia drops from 1000 to 300), then smooth decrease (<i>K</i> = 5 : 280, <i>K</i> = 6: 265). Elbow point at <i>K</i> = 4 – optimal number of clusters for data
СНИЖЕНИЕ РАЗМЕРНОСТИ	Process of reducing number of features (variables) in data while preserving important information and structure. Dimensionality reduction helps visualize high-dimensional data, reduce computational costs, and	Original data contains 100 features about customers (demographics, behavior, preferences). After PCA (Principal Component Analysis) data is projected to 3 principal components preserving 85 %

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	avoid curse of dimensionality	variance. Now data can be visualized in 3D and models can be trained faster
МЕТОД ГЛАВНЫХ КОМПОНЕНТ	Dimensionality reduction method that transforms original correlated features into new set of uncorrelated features (principal components), ordered by amount of explained variance. First component explains largest variance, second – next largest, etc.	Analysis of anthropometric data (height, weight, chest, waist, hips, arm and leg length – 15 features). PCA shows: first principal component (overall body size) explains 60 % variance, second (proportions) – 20 %, third – 10 %. First 3 components suffice instead of 15 original features
<i>T-SNE</i>	Method for visualizing high-dimensional data by projecting it to 2D- or 3D-space while preserving local structure (close points in original space remain close). <i>t</i> -SNE is particularly good for visualizing clusters	Visualization of 10,000 MNIST handwritten digit images (each image – 784 features). <i>t</i> -SNE projects data to 2D: the graph clearly shows 10 separate clusters corresponding to digits 0–9, though original data was 784-dimensional
UMAP	Dimensionality reduction method preserving both local and global data structure. UMAP is faster than <i>t</i> -SNE and scales better to large datasets, while often preserving more global structure	Analysis of genetic data of 50,000 people (20,000 genes). UMAP projects data to 2D in 10 minutes (<i>t</i> -SNE would take several hours). Map shows clusters by ethnic origin (Europeans, Asians, Africans) and gradients within groups
СОБСТВЕННОЕ ЗНАЧЕНИЕ	Characteristic number of covariance matrix in PCA showing amount of data variance captured by corresponding principal	Covariance matrix of 15 features has eigenvalues: $\lambda_1 = 8,5$; $\lambda_2 = 3,2$; $\lambda_3 = 1,8$; $\lambda_4 = 0,9...$ First

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	component. Larger eigenvalue means more information (variability) component preserves	principal component ($\lambda_1 = 8,5$) explains $8,5/15 = 57\%$ of total variance, first three components together – $(8,5 + 3,2 + 1,8)/15 = 90\%$
СОБСТВЕННЫЙ ВЕКТОР	Vector defining direction of principal component in feature space; shows how original features combine to form new component. Eigenvector is normalized (unit length)	First eigenvector of anthropometric data PCA: [0,35; 0,36; 0,34; 0,35; 0,33; ...] – all coefficients positive and approximately equal, so first component is “overall body size” (averaging all measurements). Second vector: [0,1; -0,5; 0,2; -0,6; 0,3; ...] – alternating signs show “proportions”
ОБЪЯСНЕННАЯ ДИСПЕРСИЯ	Proportion of total data variance preserved by component or set of components. Explained Variance Ratio shows percentage of variance explained by each principal component	PCA of 20 features: first component explains 45 % variance, second – 20 %, third – 12 %, fourth – 8 %. First 4 components together explain 85 % of total data variability – remaining 16 components can be dropped with only 15 % information loss
АССОЦИАТИВНЫЕ ПРАВИЛА	Rules describing relationships between variables (items) in form “if X , then Y ”. Association rules are used for market basket analysis and discovering frequently co-occurring items	Supermarket receipt analysis reveals rule: {beer, chips} → {peanuts} with support 5 %, confidence 70 %, and lift 3,2. This means: 5 % of all purchases contain all three items; of buyers who took beer and chips, 70 % also take peanuts

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ПОДДЕРЖКА	Frequency of itemset appearance in data; proportion of transactions (observations) containing given itemset. Support = $P(X \cup Y)$ – probability of joint appearance of items	In database of 10,000 purchases: 500 contain {milk, bread}, support = $500/10\,000 = 0,05$ (5 %). Rule {milk} → {bread} has support 5 % – this is relatively frequent combination of goods
ДОСТОВЕРНОСТЬ	Conditional probability of item Y appearing given item X is present; shows how often rule is true. Confidence = $P(Y X) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$	Out of 1000 purchases: 200 contain milk, 150 contain {milk, bread}. Confidence ({milk} → {bread}) = $150/200 = 0,75$ (75 %). This means: 75 % of buyers who took milk also take bread
ЛИФТ	Measure of association rule strength showing how often items occur together compared to random coincidence. $\text{Lift} = \frac{\text{Confidence}}{P(Y)} = \frac{P(Y X)}{P(Y)}$ Lift > 1 indicates positive relationship, Lift < 1 – negative, Lift = 1 – independence	Rule {Python online course} → {Data Science online course} has Lift = 4,5. This means: buyers, who have chosen Python online course, buy Data Science course 4,5 times more often than random buyers. Strong association despite low support (only 2 % of purchases)
ОБНАРУЖЕНИЕ АНОМАЛИЙ	Identification of non-standard observations (outliers) that differ significantly from majority of data. Anomaly detection is used for fraud detection, equipment failures, cyberattacks	Bank transaction monitoring system: typical client operations – purchases 500–5000 rubles in Moscow. Suddenly appears transaction: cash withdrawal 150,000 rubles in Bangkok. Anomaly detection algorithm (Isolation Forest) marks operation as suspicious, bank blocks card, and contacts their clients

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
МОДЕЛИ ГАУССОВСКОЙ СМЕСИ	Probabilistic clustering model assuming data is generated by mixture of several Gaussian (normal) distributions. Unlike K-means (hard clustering), GMM gives soft clustering – probability of belonging to each cluster	Voice recording clustering of 3 speakers: GMM models each voice as multivariate normal distribution in acoustic feature space. For new recording model outputs: probability of belonging to speaker 1 = 0,75; speaker 2 = 0,15; speaker 3 = 0,10
ЕМ-АЛГОРИТМ	Iterative algorithm for estimating parameters of models with hidden variables (e.g., GMM). EM alternates two steps: <i>E</i> -step (estimate probabilities of cluster membership at current parameters) and <i>M</i> -step (recalculate parameters to maximize likelihood)	Training GMM with 4 components: at <i>E</i> -step we calculate for each point probabilities of belonging to 4 clusters; at <i>M</i> -step we recalculate means, covariances, and component weights based on these probabilities. Repeat 50 iterations until convergence – parameters are stabilizing

4.2. Статистика в современных системах искусственного интеллекта и новые тенденции

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ГЛУБОКОЕ ОБУЧЕНИЕ	Subfield of machine learning using neural networks with many layers (deep neural networks). Deep learning automatically extracts hierarchical features from data: from simple (edges, textures) to complex (objects, concepts)	Object recognition in images: convolutional neural network (CNN) with 50 layers trains on millions of labeled photos. First layers learn to recognize edges and corners, middle – textures and simple shapes, deep – object parts (eyes, wheels), last – whole objects (faces, cars)

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
НЕЙРОННАЯ СЕТЬ	Computational system inspired by biological neural networks of brain; consists of interconnected nodes (neurons) organized in layers. Each neuron receives inputs, applies weighted sum and activation function, passes result further	Review sentiment classification: input layer receives review word vector (1000 features), hidden layer of 128 neurons processes information, output layer of 2 neurons outputs probabilities “positive” / “negative”. Network trains on 10,000 labeled reviews
ОБРАТНОЕ РАСПРОСТРАНЕНИЕ ОШИБКИ	Algorithm for training neural networks computing gradient of loss function with respect to all network weights using chain rule. Backpropagation propagates error from output layer to input, showing how to change each weight to reduce error	After forward pass through network (prediction), error (loss) is calculated. Back-propagation goes backward: computes gradient for output layer weights, then hidden layer, then input. Gradient descent updates weights: $w_{new} = w_{old} - learning_rate \times gradient$
ГРАДИЕНТНЫЙ СПУСК	Iterative optimization method minimizing loss function by moving in direction opposite to gradient (steepest descent). At each iteration parameters are updated: $\theta = \theta - \alpha \times \nabla J(\theta)$, where α is learning rate, ∇J is loss gradient	Linear regression training: initial error $MSE = 100$. Gradient descent with $learning_rate = 0,01$ makes 1000 steps: after 100 iterations $MSE = 45$; after 500 iterations $MSE = 12$; after 1000 iterations $MSE = 8,5$ (convergence)
СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ СПУСК	Variant of gradient descent using random samples (mini-batches) instead of entire dataset to compute	Training neural network on 1 million images: full gradient descent would require pass through all 1M images for one update (very slow). SGD uses mini-batches of 3

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	gradient. SGD updates parameters faster and can escape local minima due to noise, but trajectory is less stable	2 images: after each 32 it updates weights. In one epoch 31,250 updates occur instead of one
РАЗМЕР БАТЧА	Number of examples processed in one step (before weight update). Batch Size affects memory requirements, training stability, and convergence speed. Small batch – noisy updates but fast convergence; large batch – stable gradients but slower and requires more memory	Experiments with batch size for CNN training: batch _{size} = 32 – training unstable (loss jumps) but quickly finds good solution; batch _{size} = 256 – stable smooth loss decrease, optimal speed; batch _{size} = 2048 – very stable but requires much memory and converges slowly
ЭПОХА	One complete pass through entire training dataset. During one epoch model sees each training example exactly once (in SGD – as part of different batches). Training usually requires multiple epochs for convergence	Training model on 50,000 images with batch _{size} = 100: one epoch = 500 iterations (50,000/100). After 1 epoch validation accuracy 45 %; after 10 epochs – 78 %; after 30 epochs – 85 %; after 50 epochs – 86 % (saturation, further training doesn't improve)
СКОРОСТЬ ОБУЧЕНИЯ	Parameter update step in gradient descent; hyperparameter determining how strongly weights change at each iteration. Too high learning rate causes divergence (oscillation), too low – slow training or getting stuck in local minimum	Testing learning _{rate} : lr = 0,1 – loss grows, model diverges; lr = 0,01 – rapid loss decrease, convergence in 50 epochs; lr = 0,001 – very slow decrease, after 100 epochs loss still high; lr = 0,0001 – training practically doesn't proceed. Optimal lr = 0,01

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
ФУНКЦИЯ АКТИВАЦИИ	Function applied to neuron's weighted sum of inputs, introducing nonlinearity into neural network. Without nonlinear activation functions, multi-layer network would be equivalent to one linear layer. Popular functions: ReLU, Sigmoid, Tanh, Softmax	Neuron calculated $z = 2,5$ (weighted sum of inputs). ReLU is applied: $\text{output} = \max(0, 2,5) = 2,5$. For another neuron $z = -1,2$, ReLU: $\text{output} = \max(0, -1,2) = 0$ (neuron is “dead”, doesn’t activate)
RELU (Релу)	Linear activation function with rectification: $f(x) = \max(0, x)$. ReLU returns input if positive, and 0 otherwise. ReLU is most popular activation function in deep learning due to computation simplicity and solving vanishing gradient problem	Convolutional network for image recognition uses ReLU after each convolutional layer: if neuron output $-3,2$; ReLU zeroes it (0), if output $5,7$; ReLU passes it (5,7). ReLU accelerates training 6 times compared to sigmoid and tanh
SOFTMAX	Activation function for multiclass classification, transforming vector of raw scores (logits) into probability distribution where all values are positive and sum to 1. Softmax is used in output layer for tasks with $K > 2$ classes	Neural network classifies image into 3 classes (cat, dog, bird). Output layer gave logits: $[2,1; 0,5; -0,3]$. Softmax transforms: $P(\text{cat}) = 0,78$; $P(\text{dog}) = 0,16$; $P(\text{bird}) = 0,06$. Sum of probabilities = 1,0. Prediction: cat (highest probability)
СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ	Specialized neural networks for processing images and other data with grid structure. CNN use convolutional layers applying filters (kernels) to local areas of input data, automatically	MNIST handwritten digit recognition: CNN receives 28×28 pixel image. First convolutional layer applies 32 filters 3×3 , extracting simple features (edges, corners). Second layer – 64 filters, extracts more complex patterns. After

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	extracting spatial features	pooling and fully connected layers, network classifies digit with 99,2 % accuracy
РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ	Neural networks for processing sequential data (time series, text, speech) where output at previous step is passed as input at next step. RNN have “memory” of previous sequence elements	Stock price forecasting: RNN receives prices for last 30 days as input. At step $t = 1$ it processes day 1 price, saves hidden state h_1 . At step $t = 2$ it processes day 2 price together with h_1 , gets h_2 , and so on to day 30. Final hidden state h_{30} contains information about entire sequence for day 31 forecast
TRANSFORMER	Neural network architecture based on attention mechanism that revolutionized natural language processing. Transformer processes entire sequence in parallel (unlike RNN) and efficiently captures long-term dependencies	English → Russian machine translation model based on Transformer (like in “Yandex Translate”): input sentence “I love machine learning” is split into tokens, each token transformed into vector. Self-attention computes relationships between all words in parallel. Encoder encodes sentence meaning, Decoder generates translation token by token
МЕХАНИЗМ ВНИМАНИЯ	Mechanism allowing model to focus on relevant parts of input data when processing each element. Attention computes importance weights for each sequence element, allowing model to “pay attention” to needed words/features	Translating sentence “The cat sat on the mat because it was tired”. When translating word “it”, attention mechanism assigns high weights to words “cat” (0,7) and “tired” (0,2), low – to others. This helps model understand that “it” refers to cat, not mat, and translate correctly
ТРАНСФЕРНОЕ ОБУЧЕНИЕ	Using pre-trained model for solving new but related task. Instead of training from scratch, model	Chest X-ray classification into pneumonia/normal (only 5000 images). Instead of training CNN from scratch, pre-trained ResNet-50 is

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	trained on large data-base (e.g., ImageNet) is taken and fine-tuned on smaller dataset of specific task	taken (trained on ImageNet, 1,2M images). Last layers are replaced and fine-tuned on medical images. Accuracy 94 vs 78 % when training from scratch – transfer learning gave +16 %
АНСАМБЛЕВЫЕ МЕТОДЫ	Combination of several models (ensemble) to get better results than individual models. Ensembles reduce variance, bias, or both. Main approaches: bagging, boosting, stacking	Customer churn prediction: single decision tree gives 76 % accuracy, random forest (ensemble of 100 trees, bagging) – 84 %, gradient boosting (ensemble of 200 weak trees, boosting) – 87 %. Ensembles significantly outperform single models
БЭГГИНГ	Ensemble method training multiple models in parallel on different bootstrap samples (random subsamples with replacement) from training data. Final prediction is averaging (regression) or voting (classification). Bagging reduces variance	Random Forest is classic example of bagging: 100 decision trees are trained. Each tree trains on its bootstrap sample (random 80 % of data with repetitions). When predicting, all 100 trees vote, final class is majority vote. Accuracy 89 vs 76 % for single tree
БУСТИНГ	Ensemble method with sequential training where each next model focuses on errors of previous ones. Weak learners are combined into strong learner. Popular algorithms: AdaBoost, Gradient Boosting, XGBoost, LightGBM	Spam classification: first tree correctly classifies 70 % of emails. Second tree trains on errors of first (more weight to misclassified emails), accuracy on difficult cases 65 %. Third tree corrects errors of first two. After 100 trees ensemble achieves 96 % accuracy
СЛУЧАЙНЫЙ ЛЕС	Ensemble of decision trees combining bagging (random data samples) and random feature selection at	Apartment rental price prediction: Random Forest of 200 trees considers 15 features (area, district, floor, metro, etc.). Each tree builds

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	each node. Random Forest reduces overfitting of individual trees and ensures high accuracy	on random data subsample, at each node random subset of features is chosen ($\sqrt{15} \approx 4$ features). Final accuracy $R^2 = 0,87$
МАСШТАБИРУЕМОСТЬ	System's ability to cope with growth of data volume, number of users, or computational complexity without significant performance loss. Scalability is critical for deploying ML models in production	YouTube recommendation model must process 1 billion requests per day from 2 billion users. System scales horizontally: adding new servers from 10 to 100 allows processing 10 times more requests with linear performance growth
РАСПРЕДЕЛЕННЫЕ ВЫЧИСЛЕНИЯ	Data processing on multiple nodes (computers, servers) in parallel to accelerate computations and process large data volumes that don't fit on one computer	Training model on 10 terabytes of user logs: one server with 1 TB disk can't handle. Distributed system of 50 nodes (each with 1 TB) splits data: each node processes its 200 GB in parallel. Training completes in 4 hours instead of 200 hours on one server
HADOOP	Open-source framework for distributed big data processing using MapReduce programming model. Hadoop consists of HDFS (distributed file system) and MapReduce (parallel processing)	Web-site log analysis (50 TB per month): data stored in HDFS on 100 nodes. MapReduce task: Map – count views of each page on each node, Reduce – aggregate results from all nodes. Final report on 10,000 pages generated in 30 minutes
SPARK	Fast engine for big data processing supporting iterative algorithms and interactive queries. Spark stores data in memory (unlike Hadoop MapReduce which writes to disk),	Training ML model on 1 TB data: Hadoop MapReduce requires 10 iterations, each reads data from disk – 5 hours. Apache Spark loads data to memory once, 10 iterations work with memory – 15 minutes. Speedup 20 times

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	making it 10–100 times faster for many tasks	
ОНЛАЙН-ОБУЧЕНИЕ	Incremental model training as new data arrives, without need to retrain on all data from scratch. Online learning adapts to changes in data (concept drift) and works with streaming data	Click-through rate (CTR) prediction model: every 5 minutes new data about 10,000 impressions and clicks arrives. Online algorithm (SGD) updates model weights on new data, adapting to changes in user behavior. Model is always current without full retraining
ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ	Type of machine learning where agent learns through interaction with environment: performs actions, receives rewards or penalties, and learns to maximize cumulative reward. RL is used in games, robotics, autonomous systems	Training agent to play chess: agent makes move (action), environment (game) transitions to new state (position on board), agent receives reward (+1 for win, –1 for loss, 0 for draw). After millions of games agent learns strategy defeating grandmasters (AlphaZero)
ГЕНЕРАТИВНЫЙ ИИ	Artificial intelligence capable of creating new content (text, images, music, code) resembling human-created content. Generative models (GPT, DALL-E, Stable Diffusion) train on large data and generate new examples from learned distribution	ChatGPT generates answer to user's question: model analyzes dialogue context, predicts next word, then next, creating coherent meaningful text. Midjourney creates image from description “astronaut riding horse in cyberpunk style” – image that didn't exist before generation
БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ	Natural language processing models with billions of parameters trained on huge text corpora. LLMs (GPT-4, PaLM, LLaMA) can generate	GPT-4 with 1,7 trillion parameters is trained on hundreds of billions of words from internet, books, articles. Model can: write essay about AI impact on economy, solve math

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	text, answer questions, translate, write code, and perform many tasks without special fine-tuning	problem with explanation, translate text from Chinese to Swahili, debug Python code – all without additional training
ЭТИЧНЫЙ ИИ	Development and use of artificial intelligence considering ethical principles: fairness, transparency, accountability, privacy, security. Ethical AI prevents harm and discrimination	Credit scoring system is checked for ethics: model should not discriminate by gender, age, race (fairness). Decisions are explainable – can understand why application was rejected (transparency). In case of error, decision can be appealed (accountability). Client data is protected (privacy)
СПРАВЕДЛИВОСТЬ	Absence of discrimination in algorithms; requirement that AI systems do not reproduce or amplify biases against certain groups of people (by gender, race, age, etc.)	Hiring algorithm analyzes resumes: testing shows model rejects women for technical positions 2 times more often than men with similar qualifications (bias). After correction (removing gender features, rebalancing data) difference reduces to 5 % – model became fairer
ОБЪЯСНИМОСТЬ	Ability to understand and interpret model decisions; capability to explain why model made particular decision. Explainability is critical for critical applications (medicine, finance, law) and building trust in AI	Model denied credit. Explainability system (SHAP) shows: main reasons for denial – high debt burden (contribution 45 %), short credit history (contribution 30 %), low income (contribution 25 %). Client understands reasons and can fix situation (increase income, pay off debts)
ИНТЕРПРЕТИРУЕМОСТЬ	Degree of understanding cause-effect relationships in model; how much human can understand logic of model decision-making. Interpretable	Linear regression predicts apartment price: $\text{price} = 50,000 \times \text{area} + 500,000 \times (\text{proximity to metro}) - 200,000 \times \text{floor}$. Interpretation is obvious: each square meter adds 50k rubles,

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	models (linear regression, decision trees) are more transparent than “black boxes” (deep neural networks)	proximity to metro adds 500k, each floor above first decreases price by 200k
ДРЕЙФ МОДЕЛИ	Change in model performance over time due to change in statistical properties of data or target variable. Model drift requires monitoring and retraining models	Demand prediction model trained on 2019 data. In 2020 pandemic started, buyer behavior changed: demand for masks and antiseptics grew 100 times, for office clothes fell. Model not retrained on new data gives huge forecast errors – concept drift occurred
MLOPS	Practice of deploying and maintaining machine learning models in production, combining machine learning, DevOps, and data engineering. MLOps includes: training automation, model monitoring, versioning, CI/CD for ML	Company deployed 50 ML models in production. MLOps platform automatically: tracks model quality (drift monitoring), retrains models on new data every week (automation), tests new versions before deployment (CI/CD), rolls back model if quality drops (safety)
КОНСТРУИРОВАНИЕ ПРИЗНАКОВ	Process of creating, selecting, and transforming features (features) from raw data to improve machine learning model performance. Feature Engineering often impacts result more than algorithm choice	Original data: transaction date, amount, store. Feature Engineering creates new features: day of week (Monday = 0, ..., Sunday = 6), time of day (morning/day/evening/night), day of month (beginning/middle/end), amount relative to client’s average check (abnormally large/small), time since last purchase. New features increase fraud detection model accuracy from 78 to 91 %
AUTOML	Automation of machine learning process: automatic model selection,	Business analyst without ML experience uploads sales data (100,000 rows, 20 features) to AutoML

ТЕРМИН	ЗНАЧЕНИЕ	ПРИМЕР
	hyperparameter tuning, feature engineering, validation. AutoML allows non-experts to create effective ML models without deep knowledge	platform (Google AutoML, H2O.ai). System automatically: tries 15 algorithms (from linear regression to gradient boosting), tunes hyperparameters for each, selects best model (XGBoost with 92 % accuracy), creates API for predictions. Whole process took 2 hours without manual intervention

Указатель терминов

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
АПРИОРНАЯ ВЕРОЯТНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
АССОЦИАТИВНЫЕ ПРАВИЛА		Раздел 4. Продвинутое тем: методы без учителя, масштабирование и интеграция искусственного интеллекта
АПОСТЕРИОРНАЯ ВЕРОЯТНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
АСИММЕТРИЯ РАСПРЕДЕЛЕНИЯ		Раздел 2. Разведочный анализ данных и описательная статистика
АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА		Раздел 2. Разведочный анализ данных и описательная статистика
АНСАМБЛЕВЫЕ МЕТОДЫ		Раздел 4. Продвинутое тем: методы без учителя, масштабирование и интеграция искусственного интеллекта
ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ		Раздел 4. Продвинутое тем: методы без учителя, масштабирование и интеграция искусственного интеллекта
ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ		Раздел 4. Продвинутое тем: методы без учителя, масштабирование и интеграция искусственного интеллекта
ОБУЧЕНИЕ С УЧИТЕЛЕМ		Раздел 3. Статистические модели для обучения с учителем
ОБЪЕМ ДАННЫХ		Раздел 1. Основы теории вероятностей и статистического мышления
ОБРАТНОЕ РАСПРОСТРАНЕНИЕ ОШИБКИ		Раздел 4. Продвинутое тем: методы без учителя, масштабирование и интеграция искусственного интеллекта
ОЧИСТКА ДАННЫХ		Раздел 2. Разведочный анализ данных и описательная статистика
ОСТАТОК		Раздел 3. Статистические модели для обучения с учителем
ОПИСАТЕЛЬНАЯ СТАТИСТИКА		Раздел 2. Разведочный анализ данных и описательная статистика

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ОНЛАЙН-ОБУЧЕНИЕ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
ОШИБКА ПЕРВОГО РОДА		Раздел 2. Разведочный анализ данных и описательная статистика
ОШИБКА ВТОРОГО РОДА		Раздел 2. Разведочный анализ дан- ных и описательная статистика
БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
БЭГГИНГ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
БИНОМИАЛЬНОЕ РАСПРЕ- ДЕЛЕНИЕ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
БОЛЬШИЕ ДАННЫЕ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
БУСТИНГ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
БИНАРНАЯ КЛАССИФИКАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
ВАЛИДАЦИОННАЯ ВЫБОРКА		Раздел 3. Статистические модели для обучения с учителем
ВИЗУАЛИЗАЦИЯ ДАННЫХ		Раздел 2. Разведочный анализ данных и описательная статистика
ВЫБРОС		Раздел 2. Разведочный анализ данных и описательная статистика
ГЕНЕРАТИВНЫЙ ИИ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
ГЕТЕРОСКЕДАСТИЧНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
ГИПЕРПАРАМЕТРЫ		Раздел 3. Статистические модели для обучения с учителем
ГИСТОГРАММА		Раздел 2. Разведочный анализ данных и описательная статистика

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ГЛУБОКОЕ ОБУЧЕНИЕ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ГОМОСКЕДАСТИЧНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
ГРАДИЕНТНЫЙ СПУСК		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ГРАНИЦА ПРИНЯТИЯ РЕШЕНИЯ		Раздел 3. Статистические модели для обучения с учителем
ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ		Раздел 2. Разведочный анализ данных и описательная статистика
ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ		Раздел 2. Разведочный анализ данных и описательная статистика
ДОСТОВЕРНОСТЬ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ДОЛЯ ИСТИННО ПОЛОЖИТЕЛЬНЫХ		Раздел 3. Статистические модели для обучения с учителем
ДОЛЯ ЛОЖНО-ПОЛОЖИТЕЛЬНЫХ		Раздел 3. Статистические модели для обучения с учителем
ДИСКРЕТНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА		Раздел 1. Основы теории вероятностей и статистического мышления
ДИСПЕРСИОННЫЙ АНАЛИЗ		Раздел 2. Разведочный анализ данных и описательная статистика
ДИСПЕРСИЯ		Раздел 1. Основы теории вероятностей и статистического мышления
ДИАГРАММА РАССЕЯНИЯ		Раздел 2. Разведочный анализ данных и описательная статистика
ДЕНДРОГРАММА		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
DBSCAN		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ЕМ-АЛГОРИТМ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
ЗАКОН БОЛЬШИХ ЧИСЕЛ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
ЗАВИСИМАЯ ПЕРЕМЕН- НАЯ		Раздел 3. Статистические модели для обучения с учителем
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
ИСТИННО ОТРИЦАТЕЛЬНЫЙ		Раздел 3. Статистические модели для обучения с учителем
ИСТИННО ПОЛОЖИТЕЛЬНЫЙ		Раздел 3. Статистические модели для обучения с учителем
ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ		Раздел 2. Разведочный анализ данных и описательная статистика
ИНТЕРПРЕТИРУЕМОСТЬ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
КВАНТИЛИ		Раздел 2. Разведочный анализ данных и описательная статистика
КОВАРИАЦИЯ		Раздел 2. Разведочный анализ дан- ных и описательная статистика
КОМПРОМИСС СМЕЩЕНИЯ И ДИСПЕР- СИИ		Раздел 3. Статистические модели для обучения с учителем
КОНСТРУИРОВАНИЕ ПРИЗНАКОВ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
КРОСС-ВАЛИДАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
К-КРАТНАЯ КРОСС- ВАЛИДАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
КРИТИЧЕСКАЯ ОБЛАСТЬ		Раздел 2. Разведочный анализ данных и описательная статистика

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
КЛАСТЕРИЗАЦИЯ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
КЛАССИФИКАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ		Раздел 3. Статистические модели для обучения с учителем
КОЭФФИЦИЕНТЫ РЕГРЕССИИ		Раздел 3. Статистические модели для обучения с учителем
КОРРЕЛЯЦИЯ		Раздел 2. Разведочный анализ данных и описательная статистика
КОРРЕЛЯЦИЯ ПИРСОНА		Раздел 2. Разведочный анализ данных и описательная статистика
КОРРЕЛЯЦИЯ СПИРМЕНА		Раздел 2. Разведочный анализ данных и описательная статистика
ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА		Раздел 1. Основы теории вероятностей и статистического мышления
ЛАССО-РЕГРЕССИЯ		Раздел 3. Статистические модели для обучения с учителем
ЛИФТ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ЛИНЕЙНАЯ РЕГРЕССИЯ		Раздел 3. Статистические модели для обучения с учителем
ЛОЖНООТРИЦАТЕЛЬНЫЙ		Раздел 3. Статистические модели для обучения с учителем
ЛОЖНОПОЛОЖИТЕЛЬНЫЙ		Раздел 3. Статистические модели для обучения с учителем
МАШИННОЕ ОБУЧЕНИЕ		Раздел 1. Основы теории вероятностей и статистического мышления
МАРГИНАЛЬНАЯ ВЕРОЯТНОСТЬ		Раздел 1. Основы теории вероятностей и статистического мышления
МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ		Раздел 1. Основы теории вероятностей и статистического мышления
МАТРИЦА ОШИБОК		Раздел 3. Статистические модели для обучения с учителем

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ		Раздел 2. Разведочный анализ данных и описательная статистика
МЕРЫ РАЗБРОСА		Раздел 2. Разведочный анализ данных и описательная статистика
МЕЖКВАРТИЛЬНЫЙ РАЗМАХ		Раздел 2. Разведочный анализ данных и описательная статистика
МЕТОД ГЛАВНЫХ КОМПОНЕНТ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
МЕТОД ЛОКТЯ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
МОДА		Раздел 2. Разведочный анализ данных и описательная статистика
ДРЕЙФ МОДЕЛИ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ		Раздел 3. Статистические модели для обучения с учителем
МОЩНОСТЬ ТЕСТА		Раздел 2. Разведочный анализ данных и описательная статистика
МУЛЬТИКОЛЛИНЕАР- НОСТЬ		Раздел 3. Статистические модели для обучения с учителем
НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР		Раздел 3. Статистические модели для обучения с учителем
НАУКА О ДАННЫХ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
НЕЗАВИСИМАЯ ПЕРЕМЕННАЯ		Раздел 3. Статистические модели для обучения с учителем
НЕЗАВИСИМОСТЬ СОБЫТИЙ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
НЕДООБУЧЕНИЕ		Раздел 3. Статистические модели для обучения с учителем
НЕЙРОННАЯ СЕТЬ		Раздел 4. Продвинутое темы: методы без учителя,

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
		масштабируемость и интеграция искусственного интеллекта
НЕПРЕРЫВНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
НУЛЕВАЯ ГИПОТЕЗА		Раздел 2. Разведочный анализ данных и описательная статистика
ОБЪЯСНЕННАЯ ДИСПЕРСИЯ		Раздел 4. Продвинутое тематическое методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
ОБЪЯСНИМОСТЬ		Раздел 4. Продвинутое тематическое методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
<i>P</i> -ЗНАЧЕНИЕ		Раздел 2. Разведочный анализ данных и описательная статистика
ПЕРЕОБУЧЕНИЕ		Раздел 3. Статистические модели для обучения с учителем
ПЕРЦЕНТИЛИ		Раздел 2. Разведочный анализ данных и описательная статистика
ПОДДЕРЖКА		Раздел 4. Продвинутое тематическое методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ		Раздел 3. Статистические модели для обучения с учителем
ПОЛНОТА		Раздел 3. Статистические модели для обучения с учителем
ПРАВДОПОДОБИЕ		Раздел 3. Статистические модели для обучения с учителем
ПРЕДИКТИВНАЯ АНАЛИТИКА		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
ПРЕЦИЗИОННОСТЬ		Раздел 3. Статистические модели для обучения с учителем
ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ ИСХО- ДОВ		Раздел 1. Основы теории вероят- ностей и статистического мышле- ния
ПРОПУЩЕННЫЕ ДАННЫЕ		Раздел 2. Разведочный анализ данных и описательная статистика
ПРОСТАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ		Раздел 3. Статистические модели для обучения с учителем

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ		Раздел 2. Разведочный анализ данных и описательная статистика
РАСПРЕДЕЛЕНИЕ БЕРНУЛЛИ		Раздел 1. Основы теории вероятно- стей и статистического мышления
РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ		Раздел 1. Основы теории вероятно- стей и статистического мышления
РАСПРЕДЕЛЕНИЕ ПУАССОНА		Раздел 1. Основы теории вероятно- стей и статистического мышления
РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ		Раздел 2. Разведочный анализ данных и описательная статистика
РАЗМАХ		Раздел 2. Разведочный анализ данных и описательная статистика
РАЗМЕР ВЫБОРКИ		Раздел 2. Разведочный анализ данных и описательная статистика
РАЗМЕР ЭФФЕКТА		Раздел 2. Разведочный анализ данных и описательная статистика
РАЗМЕР БАТЧА		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция ИИ
РЕГУЛЯРИЗАЦИЯ		Раздел 3. Статистические модели для обучения с учителем
РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ		Раздел 4. Продвинутое темы: методы без учителя, масштабируе- мость и интеграция искусствен- ного интеллекта
РОС-КРИВАЯ		Раздел 3. Статистические модели для обучения с учителем
СВОБОДНЫЙ ЧЛЕН		Раздел 3. Статистические модели для обучения с учителем
СОБЫТИЕ		Раздел 1. Основы теории вероятно- стей и статистического мышления
СКОРОСТЬ ГЕНЕРАЦИИ ДАННЫХ		Раздел 1. Основы теории вероятно- стей и статистического мышления
СКОРРЕКТИРОВАННЫЙ R^2		Раздел 3. Статистические модели для обучения с учителем
СЛУЧАЙНАЯ ВЕЛИЧИНА		Раздел 1. Основы теории вероятно- стей и статистического мышления
СЛУЧАЙНЫЙ ЭКСПЕРИМЕНТ		Раздел 1. Основы теории вероятно- стей и статистического мышления

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
СЛУЧАЙНЫЙ ЛЕС		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
МОДЕЛЬ ГАУССОВСКОЙ СМЕСИ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
СКОРОСТЬ ОБУЧЕНИЯ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
СКОЛЬЗЯЩИЙ КОНТРОЛЬ		Раздел 3. Статистические модели для обучения с учителем
СПРАВЕДЛИВОСТЬ		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ		Раздел 2. Разведочный анализ данных и описательная статистика
СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА		Раздел 3. Статистические модели для обучения с учителем
СТАНДАРТНОЕ ОТКЛОНЕНИЕ		Раздел 1. Основы теории вероятностей и статистического мышления
СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА		Раздел 3. Статистические модели для обучения с учителем
СТАТИСТИЧЕСКОЕ ОБУЧЕНИЕ		Раздел 1. Основы теории вероятностей и статистического мышления
СТАТИСТИЧЕСКИЙ ВЫВОД		Раздел 2. Разведочный анализ данных и описательная статистика
СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ СПУСК		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
T-SNE		Раздел 4. Продвинутое темы: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ТЕСТОВАЯ СТАТИСТИКА		Раздел 2. Разведочный анализ данных и описательная статистика

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ТОЧЕЧНОЕ ОЦЕНИВАНИЕ		Раздел 2. Разведочный анализ данных и описательная статистика
ТОЧНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
ТОЧНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
TRANSFORMER		Раздел 4. Продвинутое тем: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ТРАНСФЕРНОЕ ОБУЧЕНИЕ		Раздел 4. Продвинутое тем: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ТЕОРЕМА БАЙЕСА		Раздел 1. Основы теории вероятностей и статистического мышления
УРОВЕНЬ ЗНАЧИМОСТИ		Раздел 2. Разведочный анализ данных и описательная статистика
UMAP		Раздел 4. Продвинутое тем: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ФУНКЦИЯ АКТИВАЦИИ		Раздел 4. Продвинутое тем: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ФУНКЦИЯ ВЕРОЯТНОСТИ		Раздел 1. Основы теории вероятностей и статистического мышления
ФУНКЦИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ		Раздел 1. Основы теории вероятностей и статистического мышления
ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ		Раздел 1. Основы теории вероятностей и статистического мышления
F1-МЕРА		Раздел 3. Статистические модели для обучения с учителем
ХИ-КВАДРАТ-ТЕСТ		Раздел 2. Разведочный анализ данных и описательная статистика
ЦЕНТРОИД		Раздел 4. Продвинутое тем: методы без учителя, масштабируемость и интеграция искусственного интеллекта

ТЕРМИН	ЦВЕТ РАЗ- ДЕЛА	НАЗВАНИЕ РАЗДЕЛА
ЧУВСТВИТЕЛЬНОСТЬ		Раздел 3. Статистические модели для обучения с учителем
ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ		Раздел 2. Разведочный анализ данных и описательная статистика
ЭТИЧНЫЙ ИИ		Раздел 4. Продвинутое тематическое моделирование: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ЭПОХА		Раздел 4. Продвинутое тематическое моделирование: методы без учителя, масштабируемость и интеграция искусственного интеллекта
ЭКСЦЕСС		Раздел 2. Разведочный анализ данных и описательная статистика
ЭЛАСТИЧНАЯ СЕТЬ		Раздел 3. Статистические модели для обучения с учителем
ЯЩИК С «УСАМИ»		Раздел 2. Разведочный анализ данных и описательная статистика

Список литературы

Учебники и монографии на русском языке

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика: основы эконометрики : учебник. М. : ЮНИТИ-ДАНА, 2020. Т. 1. 432 с.
2. Воронцов К. В. Математические методы обучения по прецедентам : учеб. пособие. М. : МФТИ, 2022. 312 с.
3. Гмурман В. Е. Теория вероятностей и математическая статистика : учеб. пособие. 12-е изд., перераб. и доп. М. : Юрайт, 2023. 479 с.
4. Ефимова М. Р., Петрова О. И., Румянцев Е. В. Общая теория статистики : учебник. М. : ИНФРА-М, 2021. 416 с.
5. Кремер Н. Ш. Теория вероятностей и математическая статистика : учебник для вузов. 5-е изд. М. : ЮНИТИ-ДАНА, 2022. 575 с.
6. Магуайр С., Брюс П., Брюс Э. Практическая статистика для специалистов по данным : пер. с англ. М. : ДМК Пресс, 2021. 432 с.
7. Носко В. П. Эконометрика : учебник для вузов. М. : Дело, 2020. 464 с.
8. Ширяев А. Н. Вероятность : учеб. пособие. 4-е изд. М. : МЦНМО, 2021. 528 с.

Зарубежные учебники и статьи в научных журналах (на английском языке)

9. Bishop C. M. Pattern Recognition and Machine Learning. New York : Springer, 2006. 738 p.
10. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45, № 1. P. 5–32.
11. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge : MIT Press, 2016. 800 p.
12. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York : Springer, 2009. 745 p.
13. James G., James G., Witten D. [et al.]. An Introduction to Statistical Learning: with Applications in R. New York : Springer, 2021. 440 p.
14. Murphy K. P. Probabilistic Machine Learning: An Introduction. Cambridge : MIT Press, 2022. 800 p.
15. Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge : Cambridge University Press, 2009. 464 p.

16. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken : Pearson, 2021. 1136 p.

17. Wasserman L. All of Statistics: A Concise Course in Statistical Inference. New York : Springer, 2004. 442 p.

Статьи и материалы конференций

18. Chen T. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016. P. 785–794.

19. Vaswani A. Attention Is All You Need // Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, 2017. P. 5998–6008.

Электронные ресурсы и документация

20. Scikit-learn: Machine Learning in Python // Scikit-learn developers. URL: <https://scikit-learn.org> (дата обращения: 01.04.2026).

21. TensorFlow Documentation // Google LLC. URL: <https://www.tensorflow.org> (дата обращения: 01.04.2026).

22. PyTorch Documentation // PyTorch Foundation. URL: <https://pytorch.org> (дата обращения: 01.04.2026).

23. StatSoft Electronic Textbook // StatSoft, Inc. URL: <https://www.statsoft.com/textbook> (дата обращения: 01.04.2026).

24. Курс «Аналитик данных» // Яндекс Практикум. URL: <https://practicum.yandex.ru/data-analyst> (дата обращения: 01.04.2026).

25. Курс «Введение в машинное обучение» // Stepik. URL: <https://stepik.org/course/8057> (дата обращения: 01.04.2026).

26. Coursera. Machine Learning Specialization // Stanford Online, DeepLearning.AI. URL: <https://www.coursera.org/specializations/machine-learning-introduction> (дата обращения: 01.04.2026).

Учебное издание

СТАТИСТИЧЕСКИЕ ОСНОВЫ
АНАЛИЗА БОЛЬШИХ ДАННЫХ
=
STATISTICAL FUNDAMENTALS
OF BIG DATA ANALYSIS

Составитель

Мойко Ирина Михайловна

Корректоры: *М. Н. Кузнецова, А. А. Есавкина*

Технический редактор *С. В. Денисова*

Компьютерная верстка *С. В. Денисовой*

Подписано в печать 26.10.2026.

Формат 60×84¹/₁₆. Усл. печ. л. 8,3.

Тираж 4. Заказ № 488.

Издательство ПГУ

440026, г. Пенза, ул. Красная, 40

Тел.: (8412) 66-60-49, 66-67-77; e-mail: iic@pnzgu.ru

