Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов имени Патриса Лумумбы»

Факультет искусственного интеллекта
(наименование основного учебного подразделения (ОУП)-разработчика ОП ВО)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

СТАТИСТИЧЕСКИЕ МЕТОДЫ И ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ
(наименование дисциплины/модуля)

Рекомендована МССН для направлений подготовки:

02.03.02 ФУНДАМЕНТАЛЬНАЯ ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ;
09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА
(код и наименование направления подготовки/специальности)

Освоение дисциплины ведется в рамках реализации основной профессиональной образовательной программы высшего образования (ОП ВО):

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: РАЗРАБОТКА И ОБУЧЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ
(наименование (профиль/специализация) ОП ВО)

2026 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Статистические методы и первичный анализ данных» входит в программу бакалавриата «Искусственный интеллект: разработка и обучение интеллектуальных систем» по направлениям подготовки 02.03.02 Фундаментальная информатика и информационные технологии и 09.03.03 Прикладная информатика, и изучается в 3 семестре 2 курса. Дисциплину реализует Кафедра прикладного искусственного интеллекта. Дисциплина состоит из 3 разделов и 26 тем и направлена на изучение методов разведочного анализа данных (EDA), описательной статистики, визуализации данных, оценки качества и репрезентативности данных для машинного обучения, методов обнаружения аномалий и выбросов, анализа пропущенных значений и стратегий их обработки, корреляционного анализа и отбора признаков, предобработки и трансформации данных (нормализация, кодирование, масштабирование), инструментов экосистемы Python для работы с данными (Pandas, NumPy, Matplotlib, Seaborn, Plotly), а также практик документирования процессов работы с данными и формулирования критериев качества данных для обучения моделей ИИ.

Целью освоения дисциплины является формирование у студентов практических навыков проведения полного цикла первичного анализа данных — от сбора и профилирования до предобработки и подготовки данных для обучения моделей машинного обучения, включая способность критически оценивать качество и репрезентативность данных, выявлять смещения и аномалии, применять статистические методы для обоснования решений по обработке данных, визуализировать распределения и зависимости, документировать процессы работы с данными, а также формулировать критерии приёмки и оценки качества данных и моделей.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Статистические методы и первичный анализ данных» направлено на формирование у обучающихся следующих компетенций (части компетенций):

Таблица 2.1. Перечень компетенций, формируемых у обучающихся при освоении дисциплины (результаты освоения дисциплины)

Шифр	**Компетенция**	**Индикаторы достижения компетенции** (в рамках данной дисциплины)
УК-1	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.2 Умеет анализировать и систематизировать разнородные данные, оценивать эффективность процедур анализа проблем и принятия решений в профессиональной деятельности;
УК-12	Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку информации, ее достоверность, строить	УК-12.2 Способен проводить оценку информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
	логические умозаключения на основании поступающих информации и данных	
ОПК-1	Способен применять фундаментальные знания, полученные в области математических и естественных наук, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	ОПК-1.3 Владеет навыками проведения вычислительных экспериментов, анализа их результатов и обоснования выбора математического аппарата для решения конкретных профессиональных задач в области ИИ;
ОПК-7	Способен решать задачи профессиональной деятельности на основе информационной культуры, применяя методы сбора, обработки, анализа и интерпретации данных с использованием информационно-коммуникационных технологий	ОПК-7.1 Знает принципы организации данных, методы сбора, хранения и предобработки данных, основы информационной и библиографической культуры, требования к качеству данных для обучения моделей ИИ; ОПК-7.2 Умеет осуществлять сбор данных из различных источников, проводить разведочный анализ данных (EDA), статистический анализ, визуализацию, работать с распределёнными системами хранения и обработки данных;
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	ПК-3.1 Верифицирует требования к ПО систем ИИ, определяет требования к тестам и критерии приёмки; ПК-3.3 Оценивает результаты тестирования, реализует процесс контроля качества ПО систем ИИ;
BD-1	Способен осуществлять поиск, сбор, очистку и предварительный анализ данных	BD-1.2 Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи; BD-1.3 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ; BD-1.4 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных;
BD-2	Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных	BD-2.1 Определяет требования к наборам и качеству данных для решения задач машинного обучения;
LC-2	Способен проводить эксперименты на данных, формулировать гипотезы исследования, строить (обучать, дообучать) модели ИИ с оценкой их качества и анализом ошибок, обеспечивать воспроизводимость и масштабируемость исследований на данных	LC-2.2 Проводит эксперименты на данных и визуализирует результаты с применением технологий анализа данных (статистического анализа), методов и алгоритмов МО;
MF-1	Способен применять современную теоретическую математику для разработки	MF-1.2 Применяет аппарат теории вероятностей, математической статистики и теории информации для формулирования и анализа задач искусственного интеллекта;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
	новых алгоритмов и формулирования перспективных задач ИИ	
MF-4	Способен применять статистические методы для анализа данных, валидации моделей машинного обучения и проведения экспериментов в области ИИ	MF-4.1 Применяет статистические методы анализа и машинного обучения для решения задач анализа данных и проведения экспериментов на данных;
ML-2	Способен применять фундаментальные принципы и методы машинного обучения, включая подготовку данных, оценку качества моделей и работу с признаками	ML-2.2 Применяет методы предварительной обработки данных и работы с признаками;
PL-1	Способен применять язык программирования Python для решения задач в области ИИ	PL-1.2 Осуществляет выбор инструментов разработки на Python, приемлемых для создания прикладной системы обработки научных данных, машинного обучения и визуализации с заданными требованиями;
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	SS-3.1 Учитывает в работе когнитивные искажения человека и примеры их проявления при работе с данными и ИИ, выявляет предвзятости систем ИИ, аргументированно оценивает надежность данных и выдачи ИИ, применяет базовые принципы критического мышления (оценка источников, проверка аргументов, отличие факта от интерпретации);

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Статистические методы и первичный анализ данных» относится к обязательной части блока 1 «Дисциплины (модули)» образовательной программы высшего образования.

В рамках образовательной программы высшего образования обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Статистические методы и первичный анализ данных».

Таблица 3.1. Перечень компонентов ОП ВО, способствующих достижению запланированных результатов освоения дисциплины

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
УК-12	Способен: искать нужные источники информации и данные, воспринимать, анализировать, запоминать и передавать информацию с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач; проводить оценку	Программирование на языке Python; Введение в искусственный интеллект; Технологическая (проектно-технологическая) практика (учебная);	Методы разработки решений на основе искусственного интеллекта (Git, Docker); Введение в базы данных; *Вайб-кодинг***; Методы машинного обучения; Эксплуатационная практика (учебная);

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
	информации, ее достоверность, строить логические умозаключения на основании поступающих информации и данных		
УК-1	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	Линейная алгебра; Дискретная математика; Математический анализ; Алгоритмы и структуры данных;	Преддипломная практика; Онтология и графы знаний; Введение в базы данных; Hadoop, SPARK;
ОПК-1	Способен применять фундаментальные знания, полученные в области математических и естественных наук, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	Линейная алгебра; Дискретная математика; Математический анализ;	Дифференциальные уравнения; Методы машинного обучения; Оптимизация моделей машинного обучения; Основы глубокого обучения; Нейронные сети;
ОПК-7	Способен решать задачи профессиональной деятельности на основе информационной культуры, применяя методы сбора, обработки, анализа и интерпретации данных с использованием информационно-коммуникационных технологий	Технологическая (проектно-технологическая) практика (учебная);	Технологическая (проектно-технологическая) практика (производственная); Введение в базы данных; Онтология и графы знаний; Hadoop, SPARK; Методы машинного обучения;
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	Технологическая (проектно-технологическая) практика (учебная); Программирование на языке Python;	Преддипломная практика; Технологическая (проектно-технологическая) практика (производственная); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Методы машинного обучения; Нейронные сети; Безопасность систем искусственного интеллекта; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; Методы разработки решений на основе искусственного интеллекта (Git, Docker); MLOps и промышленная

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Оптимизация моделей машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров;
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	Правоведение; Введение в искусственный интеллект;	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (производственная); Преддипломная практика; Методы машинного обучения; Нейронные сети; Безопасность систем искусственного интеллекта; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; *Вайб-кодинг***; Оптимизация моделей машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров; Введение в компьютерное зрение; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); *Основы программирования HTML - CSS - JavaScript***; *Основы программирования на языке NodeJS***; *Основы программирования на языке Go***; *Основы программирования на языке Julia***; *Основы робототехники***; *Цифровые двойники***; *Информационный поиск***; *Рекомендательные*

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			*системы***; *Обработка сигналов***; *Анализ временных рядов***; Философия; *Большие языковые модели***;
MF-1	Способен применять современную теоретическую математику для разработки новых алгоритмов и формулирования перспективных задач ИИ	Линейная алгебра; Математический анализ;	Методы машинного обучения; Нейронные сети; Основы глубокого обучения; *Анализ временных рядов***; Эксплуатационная практика (учебная);
MF-4	Способен применять статистические методы для анализа данных, валидации моделей машинного обучения и проведения экспериментов в области ИИ		Эксплуатационная практика (производственная); Методы машинного обучения; Дифференциальные уравнения; *Обработка сигналов***; *Анализ временных рядов***; MLOps и промышленная разработка систем искусственного интеллекта;
BD-1	Способен осуществлять поиск, сбор, очистку и предварительный анализ данных		Методы машинного обучения; *Информационный поиск***; Основы глубокого обучения; Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (производственная);
BD-2	Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных	Технологическая (проектно-технологическая) практика (учебная);	Эксплуатационная практика (производственная); Методы машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров; Введение в базы данных; Методы разработки решений на основе искусственного интеллекта (Git, Docker); MLOps и промышленная разработка систем искусственного интеллекта;
ML-2	Способен применять фундаментальные принципы и методы машинного обучения, включая подготовку данных, оценку качества моделей и работу с признаками	Введение в искусственный интеллект;	Методы машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров; Основы глубокого обучения;
PL-1	Способен применять язык	Технологическая (проектно-	Технологическая (проектно-

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
	программирования Python для решения задач в области ИИ	технологическая) практика (учебная); Программирование на языке Python; Алгоритмы и структуры данных;	технологическая) практика (производственная); Эксплуатационная практика (производственная); *Вайб-кодинг***; Методы машинного обучения; Основы глубокого обучения; Параллельное и распределенное программирование; Hadoop, SPARK;
LC-2	Способен проводить эксперименты на данных, формулировать гипотезы исследования, строить (обучать, дообучать) модели ИИ с оценкой их качества и анализом ошибок, обеспечивать воспроизводимость и масштабируемость исследований на данных		Эксплуатационная практика (учебная); Методы машинного обучения; Основы глубокого обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров; *Рекомендательные системы***;

* - заполняется в соответствии с матрицей компетенций и СУП ОП ВО

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Статистические методы и первичный анализ данных» составляет «4» зачетные единицы.

Таблица 4.1. Виды учебной работы по периодам освоения образовательной программы высшего образования для очной формы обучения.

Вид учебной работы	**ВСЕГО,** ак.ч.		**Семестр(-ы)**
			3
Контактная работа, ак.ч.	*51*		*51*
Лекции (**ЛК**)	17		17
Лабораторные работы (**ЛР**)	0		0
Практические/семинарские занятия (**СЗ**)	34		34
Самостоятельная работа обучающихся, ак.ч.	*66*		*66*
Контроль (экзамен/зачет с оценкой), ак.ч.	*27*		*27*
Общая трудоемкость дисциплины	**ак.ч.**	***144***	*144*
	зач.ед.	***4***	*4*

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Таблица 5.1. Содержание дисциплины (модуля) по видам учебной работы

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
Раздел 1	Разведочный анализ данных и описательная статистика	1.1	Введение в анализ данных. Типы данных и источники	Место первичного анализа в ML-пайплайне. Типы данных: числовые (непрерывные, дискретные), категориальные (номинальные, порядковые), временные, текстовые, бинарные. Источники данных: CSV, JSON, API, базы данных, веб-скрейпинг. Форматы хранения: CSV, Parquet, Feather. Экосистема инструментов: Pandas, NumPy, SciPy, Matplotlib, Seaborn	ЛК	ОПК-7.1, BD-2.1, УК-12.2
		1.2	Описательная статистика: меры центральной тенденции и разброса	Среднее, медиана, мода: определения, свойства, чувствительность к выбросам. Квантили, перцентили, межквартильный размах (IQR). Дисперсия, стандартное отклонение, коэффициент вариации. Асимметрия (skewness) и эксцесс (kurtosis). Робастные статистики: медиана, MAD. Выбор статистики в зависимости от распределения данных	ЛК	MF-4.1, MF-1.2, ОПК-7.1
		1.3	Визуализация данных: принципы и инструменты	Принципы эффективной визуализации (Tufte, Cleveland). Типы графиков и их назначение: гистограмма, boxplot, violin plot, scatter plot, bar chart, line plot, heatmap. Выбор графика в зависимости от типа данных и вопроса. Инструменты: Matplotlib (основы API), Seaborn (статистическая визуализация), Plotly (интерактивные графики). Ошибки визуализации: манипуляция осями, вводящие в заблуждение представления	ЛК	PL-1.2, ОПК-7.2, SS-3.1
		1.4	Практикум: загрузка и профилирование данных в Pandas	Загрузка данных из CSV, JSON (pd.read_csv, pd.read_json). Первичное знакомство: head, tail, shape, dtypes, info, describe. Автоматическое профилирование: ydata-profiling (pandas-profiling). Интерпретация отчёта профилирования: распределения, пропуски, корреляции, предупреждения. Документирование результатов	СЗ	ОПК-7.2, BD-1.2, LC-2.2
		1.5	Практикум: описательная статистика в Python	Вычисление статистик в Pandas: mean, median, std, quantile, describe, value_counts. Группировка и агрегация: groupby + agg. Сравнение распределений по подгруппам. Вычисление асимметрии и эксцесса (scipy.stats.skew, kurtosis).	СЗ	MF-4.1, ОПК-7.2, BD-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				Интерпретация результатов в контексте ML: что говорят статистики о качестве данных		
		1.6	Практикум: визуализация распределений	Построение гистограмм с различным числом бинов (hist, sns.histplot). KDE-оценка плотности (sns.kdeplot). Boxplot и violin plot для сравнения распределений по группам. QQ-plot для проверки нормальности. Визуализация категориальных данных: countplot, pie chart. Интерпретация: какие паттерны видны?	СЗ	PL-1.2, MF-4.1, ОПК-7.2
		1.7	Практикум: визуализация зависимостей и многомерных данных	Scatter plot с цветовым кодированием. Тепловая карта корреляционной матрицы (sns.heatmap). Pair plot для обзора парных зависимостей (sns.pairplot). Параллельные координаты. Интерактивная визуализация в Plotly: hover, zoom, filter. Практика: визуальный EDA реального датасета	СЗ	PL-1.2, ОПК-7.2, BD-1.2
		1.8	Практикум: критическая оценка данных и выявление смещений	Анализ репрезентативности выборки: сравнение распределений с генеральной совокупностью. Выявление смещений: недопредставленность групп, исторические предвзятости, selection bias. Проверка баланса классов для задач классификации. Обсуждение: как смещения в данных приводят к предвзятости моделей. Составление чек-листа качества данных	СЗ	SS-3.1, BD-1.2, ПК-3.1
		1.9	Практикум: EDA реального датасета — сквозная задача (часть 1)	Выбор датасета из открытых источников (Kaggle, UCI, OpenML). Полный цикл EDA: загрузка → профилирование → описательная статистика → визуализация распределений и зависимостей → формулирование гипотез. Документирование в Jupyter Notebook: markdown-ячейки с выводами, структурированный отчёт	СЗ	ОПК-7.2, LC-2.2, УК-1.2
Раздел 2	Предобработка данных и конструирование признаков	2.1	Пропущенные значения и выбросы	Типы пропусков: MCAR, MAR, MNAR. Визуализация паттернов пропусков (missingno). Стратегии обработки: удаление строк/столбцов, заполнение (средним, медианой, модой, интерполяцией), индикатор пропуска. Определение выбросов: IQR-метод, z-score, Isolation Forest (обзор). Стратегии обработки выбросов: удаление, winsorization, трансформация	ЛК	BD-1.3, MF-4.1, ОПК-7.1
		2.2	Трансформация и масштабирование признаков	Масштабирование: StandardScaler (z-нормализация), MinMaxScaler, RobustScaler. Логарифмическая и степенная	ЛК	BD-1.3, BD-1.4,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				трансформации (Box-Cox, Yeo-Johnson) для нормализации распределений. Кодирование категориальных признаков: Label Encoding, One-Hot Encoding, Target Encoding, Ordinal Encoding. Обработка высококардинальных категорий. Бинаризация числовых признаков		ОПК-7.1
		2.3	Конструирование признаков и отбор	Feature engineering: создание новых признаков из существующих (полиномиальные, взаимодействия, агрегаты, временные). Domain-specific features. Отбор признаков: фильтрующие методы (корреляция, mutual information, chi-squared), обёрточные (forward/backward selection, обзор), встроенные (feature importance, L1-регуляризация, обзор). Проблема мультиколлинеарности: VIF	ЛК	BD-1.4, MF-4.1, ML-2.2
		2.4	Практикум: обработка пропущенных значений	Визуализация паттернов пропусков (missingno: matrix, bar, heatmap). Определение типа пропусков (MCAR vs. MAR). Применение различных стратегий заполнения. Сравнение влияния стратегий на распределение признака. Использование SimpleImputer и KNNImputer из scikit-learn. Документирование решений	СЗ	BD-1.3, ОПК-7.2, LC-2.2
		2.5	Практикум: обнаружение и обработка выбросов	Визуализация выбросов (boxplot, scatter). Применение IQR-метода и z-score. Сравнение стратегий: удаление, clipping (winsorization), log-трансформация. Обсуждение: когда выброс — ошибка данных, а когда — редкое, но важное наблюдение. Связь с задачей обнаружения аномалий	СЗ	BD-1.3, MF-4.1, SS-3.1
		2.6	Практикум: масштабирование и трансформация признаков	Применение StandardScaler, MinMaxScaler, RobustScaler к набору данных. Визуализация распределений до и после трансформации. Логарифмическая трансформация скошенных признаков. Box-Cox трансформация (scipy.stats.boxcox). Обсуждение: какой метод выбрать для разных моделей (линейные, деревья, нейросети)	СЗ	BD-1.3, BD-1.4, MF-4.1
		2.7	Практикум: кодирование категориальных признаков	Применение One-Hot Encoding (pd.get_dummies, OneHotEncoder). Label Encoding для порядковых признаков (OrdinalEncoder). Target Encoding (category_encoders). Обработка высококардинальных категорий: группировка редких, хеширование. Обработка новых категорий в	СЗ	BD-1.3, BD-1.4, ОПК-7.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				тестовой выборке (handle_unknown)		
		2.8	Практикум: конструирование и отбор признаков	Создание новых признаков: полиномиальные (PolynomialFeatures), взаимодействия, агрегаты по группам. Отбор по корреляции с целевой переменной. Mutual information (mutual_info_classif/regression). Удаление мультиколлинеарных признаков (VIF). Визуализация feature importance после обучения простой модели	СЗ	BD-1.4, MF-4.1, ML-2.2
		2.9	Практикум: EDA реального датасета — сквозная задача (часть 2)	Продолжение работы с датасетом из раздела 1: обработка пропусков, выбросов, масштабирование, кодирование, конструирование признаков. Формирование финального набора признаков. Разбиение на train/test (train_test_split, стратификация). Документирование всех шагов предобработки в Jupyter Notebook	СЗ	BD-1.3, BD-1.4, LC-2.2
Раздел 3	Статистический анализ, оценка качества данных и валидация	3.1	Корреляционный анализ и проверка статистических гипотез в EDA	Коэффициенты корреляции: Пирсона (линейная), Спирмена (ранговая), Кендалла (ранговая). Интерпретация и ограничения: корреляция не равна причинности. Проверка статистической значимости корреляции. Применение статистических тестов в EDA: t-тест для сравнения групп, chi-squared для категориальных данных, тест Колмогорова-Смирнова для распределений. Множественные сравнения и p-hacking	ЛК	MF-4.1, MF-1.2, SS-3.1
		3.2	Оценка качества данных и метрики моделей	Систематический подход к оценке качества данных: полнота, корректность, согласованность, актуальность, уникальность. Метрики качества моделей: accuracy, precision, recall, F1 (классификация); MSE, RMSE, MAE, R-squared (регрессия). Кросс-валидация: k-fold, стратифицированная. Обсуждение: какие метрики выбрать и почему	ЛК	ML-2.2, ПК-3.1, BD-1.2
		3.3	Воспроизводимость анализа и документирование	Воспроизводимость в Data Science: фиксация random seed, версионирование данных и кода, документирование решений. Структура отчёта по EDA: цель, данные, методология, результаты, выводы, ограничения. Jupyter Notebook как инструмент воспроизводимого анализа: markdown, код, визуализации. Datasheet for Datasets: шаблон и практика заполнения	ЛК	LC-2.2, ОПК-7.1, ОПК-1.3

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
		3.4	Практикум: корреляционный анализ и статистические тесты	Построение и интерпретация матрицы корреляций (Пирсон, Спирмен). Визуализация (heatmap с аннотациями). Проверка гипотез: t-тест для сравнения признака в двух группах, chi-squared тест для независимости категориальных признаков. Интерпретация p-value. Обсуждение ошибок: путаница статистической и практической значимости	СЗ	MF-4.1, MF-1.2, SS-3.1
		3.5	Практикум: оценка качества данных и формулирование критериев приёмки	Разработка чек-листа качества данных для конкретного ML-проекта: допустимая доля пропусков, требования к балансу классов, ограничения на выбросы, требования к объёму. Формулирование критериев приёмки: «данные пригодны для обучения, если...». Автоматизация проверок (assert, great_expectations, обзор)	СЗ	ПК-3.1, BD-1.2, ОПК-7.1
		3.6	Практикум: базовое моделирование как инструмент EDA	Обучение простой модели (LogisticRegression / DecisionTreeClassifier) на подготовленных данных. Оценка метрик (accuracy, F1, confusion matrix). Feature importance. Анализ ошибок модели: на каких примерах ошибается? Связь качества данных с качеством модели. Обсуждение: что улучшить в данных?	СЗ	ML-2.2, ОПК-1.3, ПК-3.3
		3.7	Практикум: A/B-тестирование — дизайн и анализ	Постановка задачи A/B-теста: гипотеза, метрика, размер выборки (power analysis). Проведение теста на синтетических данных. Анализ результатов: t-тест, доверительный интервал. Типичные ошибки: подглядывание в результаты (peeking), множественные сравнения, неверный расчёт размера выборки. Связь с оценкой качества ML-моделей	СЗ	MF-4.1, ПК-3.1, ПК-3.3
		3.8	Практикум: итоговый проект — полный отчёт EDA	Завершение сквозной задачи: финальный отчёт по EDA реального датасета. Структура: описание данных (Datasheet), EDA с визуализациями, описательная статистика, обработка пропусков и выбросов, предобработка, конструирование признаков, базовое моделирование, выводы и рекомендации. Презентация результатов. Взаимное ревью отчётов	СЗ	ОПК-7.2, LC-2.2, УК-1.2, ОПК-1.3

* - заполняется только по **ОЧНОЙ** форме обучения: *ЛК – лекции; ЛР – лабораторные работы; СЗ – практические/семинарские занятия.*

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Таблица 6.1. Материально-техническое обеспечение дисциплины

Тип аудитории	**Оснащение аудитории**	**Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины (при необходимости)**
Лекционная	Аудитория для проведения занятий лекционного типа, оснащенная комплектом специализированной мебели; доской (экраном) и техническими средствами мультимедиа презентаций.	
Семинарская	Аудитория для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная комплектом специализированной мебели и техническими средствами мультимедиа презентаций.	Персональные компьютеры, необходимое ПО
Для самостоятельной работы	Аудитория для самостоятельной работы обучающихся (может использоваться для проведения семинарских занятий и консультаций), оснащенная комплектом специализированной мебели и компьютерами с доступом в ЭИОС.	Персональные компьютеры, необходимое ПО

* - аудитория для самостоятельной работы обучающихся указывается **ОБЯЗАТЕЛЬНО**!

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература:

1. Кулаичев, А. П. Методы и средства комплексного статистического анализа данных: учебное пособие / А.П. Кулаичев. — 5-е изд., перераб. и доп. — Москва: ИНФРА-М, 2025. — 484 с. — (Высшее образование). — DOI 10.12737/25093. - ISBN 978-5-16-020053-8. - Текст: электронный. - URL: https://znanium.ru/catalog/product/2155997

2. Григорьев, А. А. Методы и алгоритмы обработки данных: учебное пособие / А.А. Григорьев, Е.А. Исаев. — 2-е изд., перераб. и доп. — Москва: ИНФРА-М, 2024. — 383 с. + Доп. материалы [Электронный ресурс]. — (Высшее образование: Бакалавриат). — DOI 10.12737/1032305. - ISBN 978-5-16-015581-4. - Текст: электронный. - URL: https://znanium.ru/catalog/product/2084190

Дополнительная литература:

1. Тихомиров, Д. А. Статистический анализ данных. Практический курс в SPSS и Jamovi : учебник для вузов / Д. А. Тихомиров, А. Н. Пинчук. — Москва : Издательство Юрайт, 2026. — 353 с. — (Высшее образование). — ISBN 978-5-534-19186-8. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/589652

2. Дайзенрот, М. П., Фейзал, А. А., Он, Ч. С. Математика в машинном обучении = Mathematics for machine learning : докопайся до сути / М. П. Дайзенрот, А. А. Фейзал, Ч. С.

Он; пер. с англ. С. Черникова. — СПб. : Питер, 2024. — 507 с. : ил. — (Для профессионалов). — ISBN 978-5-4461-1788-8

Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. ЭБС РУДН и сторонние ЭБС, к которым студенты университета имеют доступ на основании заключенных договоров

- Электронно-библиотечная система РУДН – ЭБС РУДН https://mega.rudn.ru/MegaPro/Web
- ЭБС «Университетская библиотека онлайн» http://www.biblioclub.ru
- ЭБС «Юрайт» http://www.biblio-online.ru
- ЭБС «Консультант студента» www.studentlibrary.ru
- ЭБС «Знаниум» https://znanium.ru/

2. Базы данных и поисковые системы

- Sage https://journals.sagepub.com/
- Springer Nature Link https://link.springer.com/
- Wiley Journal Database https://onlinelibrary.wiley.com/
- Наукометрическая база данных Lens.org https://www.lens.org

Учебно-методические материалы для самостоятельной работы обучающихся при освоении дисциплины/модуля:*

1. Курс лекций по дисциплине «Статистические методы и первичный анализ данных».

* - все учебно-методические материалы для самостоятельной работы обучающихся размещаются в соответствии с действующим порядком на странице дисциплины **в ТУИС**!