МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет компьютерных технологий и прикладной математики

УТВЕРЖДАЮ:
Проректор по учебной работе,
качеству образования – первый
проректор
_____Хагуров Т.А.
подпись
« 29 » августа 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Б1.В.ДВ.01.01 «Подготовка данных машинного обучения»

Направление подготовки 01.03.02 Прикладная математика и информатика

Направленность (профиль) Современные методы машинного обучения и компьютерного зрения

Форма обучения очная

Квалификация бакалавр

Краснодар 2025

Рабочая программа дисциплины Подготовка данных машинного обучения составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 01.03.02 Прикладная математика и информатика

Программу составил(а):

Приходько Татьяна Александровна, доцент, к. т. н. _____

Ф.И.О. , должность, ученая степень, ученое звание подпись

Рабочая программа дисциплины утверждена на заседании центра искусственного интеллекта
протокол № 01 «28» августа 2025 г.
Руководитель центра ИИ Коваленко А.В.

Утверждена на заседании учебно-методической комиссии факультета Компьютерных Технологий и Прикладной Математики

протокол № 1 « 28 » августа 2025 г.

Председатель УМК факультета Коваленко А.В.

фамилия, инициалы подпись

Рецензенты:

Мостовой Евгений Викторович, генеральный директор ООО «Портал-Юг», e-mail: mostovoy@portal-yug.ru

Луценко Евгений Вениаминович, доктор экономических наук, кандидат технических наук, профессор кафедры компьютерных технологий и систем Федерального государственного бюджетное образовательное учреждение высшего образования «Кубанский государственный аграрный университет имени И.Т. Трубилина», e-mail: prof.lutsenko@gmail.com

1. Цели и задачи изучения дисциплины (модуля)

1.1 Цель освоения дисциплины «Подготовка данных машинного обучения» является формирование у студентов систематизированных знаний, практических умений и навыков применения современных методов искусственного интеллекта, машинного обучения для решения задач подготовки данных для дальнейшего применения в моделях различных предметных областей.

Дисциплина направлена на развитие способности собирать, размечать, предобрабатывать данные и оценивать качество подготовленных данных.

1.2 Задачи дисциплины

1. Кроме методов решения типовых задач подготовки данных: обработка пропущенных значений, кодирование категориальных признаков, масштабирование и нормализация числовых данных и методов обработки выбросов (аномалий) в данных, изученных ранее в дисциплине «Многомерный статистический анализ и машинное обучение», усвоить принципы и методы feature engineering (создания и преобразования признаков) для повышения эффективности моделей машинного обучения.
2. Применять на практике методы работы с несбалансированными данными и подходы к разметке данных (labeling).
3. Применять на практике критерии и метрики для оценки качества подготовленных данных и их пригодности для решения конкретной задачи.
4. Приобрести практический навык применения методов предобработки данных: очистка от шума, обработка пропусков, кодирование категориальных переменных, масштабирование признаков. Сформировать умение создавать новые признаки (Feature Engineering) на основе существующих для улучшения предсказательной способности моделей. Освоить методы селекции признаков (Feature Selection) для отбора наиболее информативных переменных и уменьшения размерности данных.
5. Приобрести умение оценивать качество подготовленного набора данных с помощью визуализации и статистических метрик перед передачей его на этап моделирования.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Подготовка данных машинного обучения» относится к части, формируемой участниками образовательных отношений Блока "Дисциплины (модули) по выбору" учебного плана (Б1.В.ДВ).

Дисциплина изучается в 6-м семестре. Для успешного освоения необходимы знания, полученные в дисциплинах: «Алгебра и введение в тензорный анализ», «Теория вероятностей и математическая статистка», «Многомерный статистический анализ», и ««Современные технологии машинного обучения»», «Программирование».

Преподавание ведется в виде лекций и лабораторных занятий с использованием интерактивных методов. Лабораторные работы направлены на практическое освоение методов и инструментов классификации на реальных данных.

Дисциплина формирует компетенции, необходимые для выполнения выпускной квалификационной работы и профессиональной деятельности в области вычислительных технологий.

1.4 Профессиональные роли в структуре образовательной программы

Роль 1: **Data Engineer (Инженер по данным)**

Задачи:

1. Проектирование и построение ETL-процессов
2. Создание и оптимизация хранилищ данных
3. Обеспечение качества и доступности данных
4. Настройка инфраструктуры для обработки больших данных
5. Интеграция разрозненных источников данных
6. Работа с данными в области природопользования, медицины, связи и телекоммуникаций

Роль 2: **ML Engineer (Инженер МО)**

Задачи:

1. Реализация ML-моделей в продуктивных системах
1. Оптимизация производительности и масштабирование моделей
1. Разработка ML-пайплайнов и автоматизация процессов
1. Мониторинг качества моделей в продуктиве
1. Интеграция ML-решений с бизнес-приложениями

Роль 3: **MLOps (Специалист по эксплуатации ИИ)**

Задачи:

1 Автоматизация процессов обучения и развертывания моделей
1. Мониторинг производительности ML-систем
1. Управление версиями моделей и данных
1. Обеспечение CI/CD для ML-проектов
1. Оптимизация вычислительных ресурсов

1.5 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих компетенций:

BD-1	**Способен осуществлять поиск, сбор, очистку и предварительный анализ данных (П)**
BD-1.1	Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи **Знает** методы заполнения пропусков в данных и удаления выбросов в табличных данных (случайные величины) Имеет навыки (**умеет**) очистки зашумленных временных рядов и изображений. Обнаруживает и устраняет выбросы в данных временных рядов. **Владеет** подходами к заполнению пропусков в данных временных рядов и изображений.
BD-1.2	Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ **Знает** основные методы понижения размерности

Умеет применить основные методы понижения размерности и подбирает оптимальную размерность в зависимости от необходимой доли объяснённой дисперсии.
Владеет методологией применения существующих библиотек, реализующих методы понижения размерности.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных
Знает и умеет применить основы методов отбора признаков и выбирает оптимальное подмножество признаков.
Владеет методологией применения существующих библиотек, реализующих методы отбора признаков.

BD-1.4 **Знает** и умеет применить методы отбора признаков.
Владеет способностью применять методы отбора признаков данных, значимых для исследования.
Умеет отбирать признаки данных, значимые для исследования,
Владеет методами finetuning

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 Знает, как сформировать требования для набора данных. Владеет умениями по формированию требований к наборам и качеству данных для решения задач машинного обучения

BD-2.2 Знает приемы и инструменты для сбора данных из разрозненных источников. Умеет работать с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность. Владеет языками и инструментами для сбора данных и оценки их корректности.

LLM-2 Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения

LLM-2.1 **Понимает принципы fine-tune**
Знает: основные подходы к тонкой настройке: полная настройка всех параметров, поэтапная разморозка слоев, методы эффективной тонкой настройки (P-Tuning, LoRA, QLoRA, Adapter). Гиперпараметры, критически важные для fine-tune: learning rate, scheduler, batch size, и их отличия от обучения с нуля.
Умеет: Отличать дообучение от первичного обучения, знает базовые процедуры **fine-tune,** анализировать задачу и выбирать наиболее подходящий метод fine-tune (полная настройка vs. эффективные методы).
Владеет: **Навыком** осознанного выбора стратегии fine-tune под ограничения (вычислительные ресурсы, объем данных, требования к качеству). Применяет fine-tune к предобученным моделям на новых датасетах.
• **Методами** анализа и интерпретации процесса дообучения (использование логов, графиков потерь).
• **Критическим мышлением** для оценки целесообразности применения fine-tune в конкретном сценарии versus использования prompt engineering или RAG.

LLM-2.2 **Создаёт обучающие наборы данных.**

Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки.Форматы данных для популярных фреймворков (Hugging Face, TensorFlow, PyTorch) и структур задач (текст-текст, текст-изображение, инстркукции и т.д.). Методы аугментации данных (data augmentation), специфичные для генеративных моделей (e.g., back-translation для текста, модификация промптов). Принципы разбиения данных на обучающую, валидационную и тестовую выборки.

Умеет: Выбирать методы с учетом требований к latency и ресурсам. собирать данные из различных источников: API, веб-скрапинг, открытые датасеты, синтетическая генерация. Очищать и предобрабатывать сырые данные: удаление шума, дубликатов, нормализация текста, приведение к единому формату. Размечать данные в соответствии с поставленной задачей (e.g., составлять пары "инструкция-ответ", аннотировать изображения). Применять методы аугментации данных для увеличения размера и разнообразия обучающего набора.

Владеет: Навыками работы с библиотеками и инструментами для обработки данных (Pandas, NumPy, Hugging Face Datasets). **Методами** обеспечения репрезентативности и сбалансированности создаваемого набора данных. **Технологиями** создания синтетических данных для задач, где реальных данных недостаточно.

Полным циклом подготовки данных: от сбора сырых данных до формирования готового для обучения объекта (DataLoader, Dataset)

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.

Знает отличия статистического обучения от не статистического, **владеет** классификацией методов статистического машинного обучения. **Умеет** применять и выбирать методы статистического машинного обучения, учитывая особенности данных и задачи, а также объясняет различия между подходами.

MF-4.2 Способен применять статистические методы для построения предсказательных моделей, включая методы для анализа и прогнозирования временных рядов, а также моделирования нестационарных случайных процессов.

Знает: теоретические основы и предположения линейной и логистической регрессии. Понятие стационарности временного ряда, методы проверки (ADF test) и приведения к стационарному виду (дифференцирование, декомпозиция). Классические модели прогнозирования временных рядов (ARIMA, SARIMA, ETS).**Умеет** формализовывать и применять статистические методы идентификации регрессионных и классификационных моделей, понимает основы базовых вероятностных моделей для временных рядов на основе авторегрессионных зависимостей.. **Владеет** приемами построения модели динамических систем для многомерных временных рядов и полей.

MF-4.3 Способен применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

Знает метрики и меры качества моделей регрессии (в т.ч. на временных рядах), классификации, кластеризации.

Умеет оценивать качество моделей МО

Владеет умением применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

ML-2 Способен применять фундаментальные принципы и методы машинного обучения включая подготовку данных оценку качества моделей и работу с признаками

ML-2.1 Различает основные типы задач машинного обучения и применяет на практике принципы их решения

Знает и различает основные типы задач машинного обучения (обучением с учителем, без учителя и с подкреплением). **Умеет** применить типовые подходы к решению базовых задач с использованием готовых инструментов и библиотек (ScikitLearn) (Б)

Умеет обоснованно применять методы решения задач машинного обучения с учётом характеристик данных и бизнес-контекста, настраивает базовые модели и проводит их оценку (П)

Владеет приемами и инструментами проектирования и реализации комплексных решений машинного обучения для нестандартных задач, включая разработку пайплайнов, оптимизацию моделей и интерпретацию результатов (Э)

Результаты обучения по дисциплине достигаются в рамках осуществления всех видов контактной и самостоятельной работы обучающихся в соответствии с утвержденным учебным планом.

Индикаторы достижения компетенций считаются сформированными при достижении соответствующих им результатов обучения.

2. Структура и содержание дисциплины

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 2 зачетных единиц (72 часа), их распределение по видам работ представлено в таблице

Виды работ	Всего часов	Форма обучения очная
		6 семестр (часы)
Контактная работа, в том числе:	**34,2**	**34,2**
Аудиторные занятия (всего):	**32**	**32**
занятия лекционного типа	16	16
лабораторные занятия	16	16
практические занятия	-	-
семинарские занятия	-	-
Иная контактная работа:	**2,2**	**2,2**
Контроль самостоятельной работы (КСР)	2	2
Промежуточная аттестация (ИКР)	0,2	0,2
Самостоятельная работа, в том числе:	**37,8**	**37,8**
Курсовая работа/проект (КР/КП) (подготовка)	-	-
Контрольная работа	-	-

Расчётно-графическая работа (РГР) (подготовка)		14	14
Выполнение индивидуальных заданий по подготовке рефератов, сообщений, презентаций		8	8
Самостоятельная проработка и материала учебников и учебных пособий, подготовка к лабораторным занятиям		9,8	9,8
Подготовка к текущему контролю		6	6
Контроль:			
Подготовка к экзамену		-	-
Общая трудоемкость	**час.**	**72**	**72**
	в том числе контактная работа	**34,2**	**34,2**
	зач. ед	**2**	**2**

2.2 Содержание дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.
Разделы/темы дисциплины, изучаемые в 6 семестре 3 курса очной формы обучения

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1.	Введение в подготовку данных для МО. Планирование датасета под задачу.	8	2		2	4
2.	Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.	8	2		2	4
3.	Разметка данных и аугментация.	8	2		2	4
4.	Предобработка изображений	8	2		2	4
5.	Предобработка аудиоданных	8	2		2	4
6.	Предобработка текстовых данных	9	2		2	5
7.	Предобработка данных для временных рядов	8,8	2		2	4,8
8.	Статистический анализ данных (EDA, визуализация, анализ распределений, корреляции, анализ признаков, feature importance).Стратегии по улучшению метрик через данные (feature engineering, отбор признаков, балансировка). Дизайн эксперимента по улучшению данных	11	2		2	7
	ИТОГО по разделам дисциплины	69,8	16		16	37,8
	Контроль самостоятельной работы (КСР)	2				
	Промежуточная аттестация (ИКР)	0,2				
	Подготовка к текущему контролю	-				
	Общая трудоемкость по дисциплине	72				

Примечание: Л – лекции, ПЗ – практические занятия / семинары, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

2.3 Содержание разделов (тем) дисциплины

2.3.1. Занятия лекционного типа

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
1.	Введение в подготовку данных для МО. Планирование датасета под задачу.	**Признаковое пространство (Feature Space).** - Какие признаки *идеально* помогут решить задачу? (e.g., для прогноза оттока: история транзакций, жалобы, активность в ленте). - Матрица "Источники данных - Потенциальные признаки". **Оценка объема данных:** Эвристики для разных типов задач (классификация, детекция, сегментация) Соотношение количества данных к сложности модели	ЛР
2.	Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.	**Сбор данных из гетерогенных источников:** CSV, базы данных (SQL), API (REST), парсинг логов. Инструменты: pandas, sqlalchemy, requests. **Создание первичного датасета:** Соединение таблиц (merge, join), аггрегация данных (например, история транзакций в признаки клиента). **Критерии качества данных на этом этапе:** **Проверка выбросов:** IQR, Z-score, Isolation Forest **Coverage (Покрытие):** Достаточно ли данных? Сколько строк/объектов? **Availability (Доступность):** Все ли запланированные признаки удалось собрать? Изображения: разрешение, освещенность, артефакты сжатия Аудио: битрейт, соотношение сигнал/шум, длительность Текст: кодировка, язык, наличие орфографических ошибок **Пример:** Планирование мультимодального датасета для задачи "распознавание эмоций" (табличные данные + аудио + текст транскрипции). **Первый взгляд на пропуски и аномалии.** **Автоматизированная проверка качества:** Скрипты для батч-проверки минимальных требований	ЛР
3.	Разметка данных и аугментация.	**Ключевые компоненты пайплайна.** **Создание пайплайна предобработки:** Обработка дубликатов, аномальных значений, приведение типов. Фиксация всех действий в коде для воспроизводимости. **Разметка данных (для задач обучения с учителем):** Определение и создание целевой переменной. Пример: Разметка клиентов на "отток" / "не отток" на основе их активности за последний месяц. Проблемы временных разрезов (Data Leakage) и как их избежать. **Продвинутая обработка аномалий:**	ЛР

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
		Алгоритмы поиска выбросов: Isolation Forest, DBSCAN Автоматический кэппинг на основе процентилей **Работа с временными рядами:** Создание лаговых признаков, скользящие статистики Извлечение признаков из даты/времени **Стратегии разметки:** Active Learning для эффективной разметки Crowdsourcing и контроль качества разметки **Пример:** Создание пайплайна предобработки для финансового датасета с временными рядами.	
4.	Предобработка изображений	**Базовые операции:** Изменение размера, кадрирование, поворот Конвертация цветовых пространств (RGB, HSV, Grayscale) **Коррекция качества:** Гистограммная эквализация, фильтрация (Гаусс, медианная). Автоконтраст, баланс белого. **Аугментация данных:** Geometric: повороты, отражения, искажения Photometric: изменение яркости, контраста, добавление шума, CutMix, MixUp - продвинутые методы Transfer Learning подготовка: Препроцессинг под конкретные архитектуры (ImageNet-нормализация) **Пример:** Написание пайплайна аугментации для задачи детекции объектов с использованием albumentations.	
5.	Предобработка аудиоданных	**Базовая обработка сигнала:** Ресемплирование, нормализация громкости Фильтрация шума, обрезка тишины **Преобразование в частотную область:** Спектрограммы (STFT), Mel-спектрограммы MFCC (Mel-frequency cepstral coefficients) - извлечение и нормализация **Аугментация для аудио:** Добавление шума, изменение pitch/tempo Time stretching, time shifting **Подготовка для моделей:** Создание дельта- и дельта-дельта features Нормализация по оси времени **Пример:** Извлечение MFCC из набора аудиофайлов и создание аугментированной версии датасета.	

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
6.	Предобработка текстовых данных	**Очистка и нормализация:** Регулярные выражения для очистки Лемматизация, стемминг, приведение к нижнему регистру **Токенизация:** Word-level, subword (BPE), character-level. Подготовка словаря. **Векторизация:** Bag of Words, TF-IDF Word2Vec, GloVe, FastText - обучение и использование **Контекстные эмбеддинги:** Подготовка данных для BERT и других трансформеров Tokenization, attention masks, padding **Пример:** Создание пайплайна от сырого текста до эмбеддингов BERT для задачи классификации тональности.	
7.	Предобработка данных для временных рядов	**Теоретические основы временных рядов.** Этапы обработки: **Работа с временной меткой (DateTime Index), Ресамплинг (Resampling) и Агрегация (включает** понижающую и повышающую дискретизацию), обработка пропусков (Missing Values), обработка выбросов и аномалий, декомпозиция, стационаризация, **Feature Engineering.**	ЛР
.	Статистический анализ данных (EDA, визуализация, анализ распределений, корреляции, анализ признаков, feature importance).Стратегии по улучшению метрик через данные (feature engineering, отбор признаков, балансировка). Дизайн	**Анализ распределений и выбросов:** Статистические тесты на нормальность (Shapiro-Wilk).Quantile-Quantile plot (Q-Q plot). **Статистический анализ связи признаков с таргетом:** **Числовые признаки vs Числовой таргет:** Коэффициент корреляции Пирсона и Спирмена, их интерпретация и ограничения. **Числовые признаки vs Категориальный таргет:** T-test и ANOVA для сравнения средних между группами. **Визуализация:** Boxplot для категориального таргета. **Анализ категориальных признаков:** Категориальный признак vs Категориальный таргет: Chi-Square test (тест хи-квадрат) на независимость. Визуализация: Stacked bar chart, heatmap. **Анализ многомерных зависимостей:** Матрица корреляций: Визуализация и интерпретация. Анализ мультиколлинеарности: Расчет VIF (Variance Inflation Factor). **Анализ для неструктурированных данных**: Изображения: анализ гистограмм, статистика по яркости/контрасту	ЛР

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
	эксперимента по улучшению данных	Аудио: анализ распределения длительностей, громкости Текст: распределение длин предложений, частотность слов **Feature Engineering на основе анализа:** Создание мета-признаков для неструктурированных данных Отбор признаков на основе статистической значимости Практика: Проведение полного статистического анализа мультимодального датасета и генерация новых признаков **Трансформации признаков на основе анализа распределения:** Логарифмирование, преобразование Бокса-Кокса для нормализации и борьбы с выбросами. **Создание новых признаков на основе доменного знания:** Пример: Из двух признаков "доход" и "количество членов семьи" создать новый "доход на члена семьи". Биннинг (дискретизация) по квантилям, по статистической значимости таргета (с помощью дерева решений). **Дисбаланс классов в задачах классификации:** Модель учится предсказывать только мажоритарный класс. Методы борьбы: Взвешивание классов (class_weight) в алгоритмах. Сэмплирование: SMOTE (синтез новых примеров меньшинства), Random Over/Under-Sampling. Метрики: Почему точность (accuracy) не всегда подходит, когда лучше применить F1-score, Precision-Recall, AUC-ROC. **Дизайн эксперимента:** A/B тестирование разных версий датасета Фиксация модели и гиперпараметров **Методы улучшения для каждого типа данных:** *Изображения*: подбор оптимальной аугментации, балансировка классов: - *Аудио*: оптимизация параметров MFCC, балансировка по длительности - *Текст*: подбор vocabulary size, оптимизация длины последовательности **Анализ ошибок и целенаправленное улучшение:** Выявление "слепых зон" модели, стратегический сбор дополнительных данных	

2.3.2. Занятия семинарского типа

Занятия семинарского типа не предусмотрены учебным планом.

2.3.3. Лабораторные работы

№	Наименование раздела (темы)	Тематика лабораторных работ	Форма текущего контроля
1.	Введение в подготовку данных для МО. Планирование датасета под задачу.	ЛР 1. Планирование и первичны сбор данных для мультимодального датасета	Опрос по теоретическому материалу. Отчет по лабораторной работе.
2.	Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.	ЛР 2. Сбор и первичная оценка мультимодальных данных - первичная обработка и очистка данных	Опрос по теоретическому материалу. Отчет по лабораторной работе.
3.	Разметка данных и аугментация.	ЛР 3. Предобработка числовых и категориальных данных	Опрос по теоретическому материалу. Отчет по лабораторной работе.
4.	Предобработка изображений	ЛР 4. Пайплайн аугментации изображений	Контрольная работа Проверка выполнения домашних работ.
5.	Предобработка аудиоданных	ЛР 5. Предобработка аудиоданных, извлечение аудиопризнаков	Опрос по теоретическому материалу. Отчет по лабораторной работе.
6.	Предобработка текстовых данных	ЛР 6. Пайплайн обработки текста	Опрос по теоретическому материалу. Отчет по лабораторной работе.
7.	Предобработка данных для временных рядов	ЛР 7. Работа с временными рядами	Опрос по теоретическому материалу. Отчет по лабораторной работе.
8.	Итоговая комплексная работа. Настройка и отбор признаков (Feature Engineering). Балансировка данных и стратегии семплирования	Итоговый проект: 1. На основе датасета из предыдущих работ (или нового) создайте новые признаки: На основе доменных знаний, путем взаимодействия существующих признаков, Используя преобразования (полиномиальные, логарифмические и т.д.). 2. Проведите отбор признаков: Методы фильтров (корреляция, взаимная информация). Методы обертки (Recursive Feature Elimination). Методы встроенные (Lasso, важность признаков в деревьях). 3. Сравните производительность модели на исходном наборе признаков и на наборе после отбора. 4. На основе несбалансированного	Опрос по теоретическому материалу. Отчет по лабораторной работе.

		датасета примените методы балансировки: Случайное недосэмплирование и пересэмплирование. SMOTE и его варианты. ADASYN. 5. Обучите модель на исходных данных и на сбалансированных и сравните метрики (Precision, Recall, F1-score, ROC-AUC).	

2.3.4 Примерная тематика курсовых работ (проектов)

Курсовая работа не предусмотрена. В качестве курсового проекта студенты защищают инфраструктуру предобработанного датасета по заданию от индустриального партнера.

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

Целью самостоятельной работы студента является:

- углубление знаний, полученных в результате аудиторных занятий;
- развитие навыков самостоятельной работы;
- закрепление опыта и знаний, полученных во время лабораторных занятий.

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	2	3
1	Проработка и повторение лекционного материала, материала учебной и научной литературы, подготовка к семинарским занятиям	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
2	Подготовка к лабораторным занятиям	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
3	Подготовка к решению задач и тестов	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
4	Подготовка к текущему контролю	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,
– в форме электронного документа,
– в форме аудио-файла,
– в печатной форме на языке Брайля.
Для лиц с нарушениями слуха:
– в печатной форме,
– в форме электронного документа.
Для лиц с нарушениями опорно-двигательного аппарата:
– в печатной форме,
– в форме электронного документа,
– в форме аудио-файла.
Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. Образовательные технологии, применяемые при освоении дисциплины (модуля)

В соответствии с требованиями ФГОС в программа дисциплины предусматривает использование в учебном процессе следующих образовательные технологии: чтение лекций с использованием мультимедийных технологий; метод малых групп, разбор практических задач и кейсов.

При обучении используются следующие образовательные технологии:

4 Технология коммуникативного обучения – направлена на формирование коммуникативной компетентности студентов, которая является базовой, необходимой для адаптации к современным условиям межкультурной коммуникации.

5 Технология разноуровневого (дифференцированного) обучения – предполагает осуществление познавательной деятельности студентов с учётом их индивидуальных способностей, возможностей и интересов, поощряя их реализовывать свой творческий потенциал. Создание и использование диагностических тестов является неотъемлемой частью данной технологии.

6 Технология модульного обучения – предусматривает деление содержания дисциплины на достаточно автономные разделы (модули), интегрированные в общий курс.

7 Информационно-коммуникационные технологии (ИКТ) - расширяют рамки образовательного процесса, повышая его практическую направленность, способствуют интенсификации самостоятельной работы учащихся и повышению познавательной активности. В рамках ИКТ выделяются 2 вида технологий:

8 Технология использования компьютерных программ – позволяет эффективно дополнить процесс обучения языку на всех уровнях.

9 Интернет-технологии – предоставляют широкие возможности для поиска информации, разработки научных проектов, ведения научных исследований.

10 Технология индивидуализации обучения – помогает реализовывать личностно-ориентированный подход, учитывая индивидуальные особенности и потребности учащихся.

11 Проектная технология – ориентирована на моделирование социального взаимодействия учащихся с целью решения задачи, которая определяется в рамках профессиональной подготовки, выделяя ту или иную предметную область.

12 Технология обучения в сотрудничестве – реализует идею взаимного обучения, осуществляя как индивидуальную, так и коллективную ответственность за решение учебных задач.

13 Игровая технология – позволяет развивать навыки рассмотрения ряда возможных способов решения проблем, активизируя мышление студентов и раскрывая личностный потенциал каждого учащегося.

14 Технология развития критического мышления – способствует формированию разносторонней личности, способной критически относиться к информации, умению отбирать информацию для решения поставленной задачи.

Комплексное использование в учебном процессе всех вышеназванных технологий стимулируют личностную, интеллектуальную активность, развивают познавательные процессы, способствуют формированию компетенций, которыми должен обладать будущий специалист.

Основные виды интерактивных образовательных технологий включают в себя:

15 работа в малых группах (команде) - совместная деятельность студентов в группе под руководством лидера, направленная на решение общей задачи путём творческого сложения результатов индивидуальной работы членов команды с делением полномочий и ответственности;

16 проектная технология - индивидуальная или коллективная деятельность по отбору, распределению и систематизации материала по определенной теме, в результате которой составляется проект;

17 анализ конкретных ситуаций - анализ реальных проблемных ситуаций, имевших место в соответствующей области профессиональной деятельности, и поиск вариантов лучших решений;

18 развитие критического мышления – образовательная деятельность, направленная на развитие у студентов разумного, рефлексивного мышления, способного выдвинуть новые идеи и увидеть новые возможности.

Подход разбора конкретных задач и ситуаций широко используется как преподавателем, так и студентами во время лекций, лабораторных занятий и анализа результатов самостоятельной работы. Это обусловлено тем, что при исследовании и решении каждой конкретной задачи имеется, как правило, несколько методов, а это требует разбора и оценки целой совокупности конкретных ситуаций.

При проведении лабораторных занятий участники закрепляют пройденный материал путем обсуждения вопросов, требующих особого внимания и понимания, отвечают на вопросы преподавателя и других слушателей, осуществляют решения тестов, направленных на повторение лекционного материала и нормативных документов по изучаемой тематике, выполняют решение задач, которые способствуют развитию практических навыков в области изучаемой дисциплины.

В число видов работы, выполняемой слушателями самостоятельно, входят:

1) поиск и изучение литературы по рассматриваемой теме;

2) поиск и анализ научных статей, монографий по рассматриваемой теме.

Интерактивные образовательные технологии, используемые в аудиторных занятиях: при реализации различных видов учебной работы (лекций и практических занятий) используются следующие образовательные технологии: дискуссии, презентации, конференции. В сочетании с внеаудиторной работой они создают дополнительные условия формирования и развития требуемых компетенций обучающихся, поскольку позволяют обеспечить активное взаимодействие всех участников. Эти методы способствуют личностно-ориентированному подходу.

Все перечисленные виды и формы учебной работы и текущего контроля направлены на формирование у обучающихся профессиональных компетенций, предусмотренных при планировании результатов обучения по дисциплине и соотнесенных с планируемыми результатами освоения образовательной программы.

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты и устанавливается особый порядок освоения указанной дисциплины. В образовательном процессе используются социально-активные и рефлексивные методы обучения, технологии социально-культурной реабилитации

с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создании комфортного психологического климата в студенческой группе.

Вышеозначенные образовательные технологии дают наиболее эффективные результаты освоения дисциплины с позиций актуализации содержания темы занятия, выработки продуктивного мышления, терминологической грамотности и компетентности обучаемого в аспекте социально направленной позиции будущего бакалавра, и мотивации к инициативному и творческому освоению учебного материала.

4. Оценочные средства для текущего контроля успеваемости и промежуточной аттестации

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины «Подготовка данных машинного обучения».

Освоение дисциплины предполагает две основные формы контроля – текущая и промежуточная аттестация.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы и предполагает овладение материалами лекций, литературы, программы, работу студентов в ходе проведения практических занятий, а также систематическое выполнение тестовых работ, решение практических задач и иных заданий для самостоятельной работы студентов. Данный вид контроля стимулирует у студентов стремление к систематической самостоятельной работе по изучению дисциплины. Он предназначен для оценки самостоятельной работы слушателей по решению задач, выполнению практических заданий, подведения итогов тестирования. Оценивается также активность и качество результатов практической работы на занятиях, участие в дискуссиях, обсуждениях и т.п. Индивидуальные и групповые самостоятельные, аудиторные, контрольные работы по всем темам дисциплины организованы единообразным образом. Для контроля освоения содержания дисциплины используются оценочные средства. Они направлены на определение степени сформированности компетенций.

Промежуточная аттестация студентов осуществляется в рамках завершения изучения дисциплины и позволяет определить качество усвоения изученного материала, предполагает контроль и управление процессом приобретения студентами необходимых знаний, умения и навыков, определяемых по ФГОС ВО по соответствующему направлению подготовки в качестве результатов освоения учебной дисциплины.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей:

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

4.1 Оценочные средства для текущего контроля успеваемости

4.1.1. Вопросы контрольного опроса в рамках занятий лекционного и семинарского типа

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины «Подготовка данных машинного обучения».

Оценочные средства включает контрольные материалы для проведения **текущего контроля** в форме тестовых заданий, кейсов и **промежуточной аттестации** в форме вопросов и заданий к **экзамену**.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Контролируемые разделы (темы) дисциплины*	Код контролируемой компетенции (или ее части)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация

1	Введение в подготовку данных для МО. Планирование датасета под задачу.	BD-1, BD-2	*Лабораторная работа №1*	*Вопросы к зачету*
2	Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.	BD-1, BD-2	*Лабораторная работа №2*	*Вопросы к зачету*
3	Разметка данных и аугментация.	BD-1, BD-2, ML-2	*Лабораторная работа №3*	*Вопросы к зачету*
4	Предобработка изображений	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №4*	*Вопросы к зачету*
5	Предобработка аудиоданных	BD-1, BD-2,	*Лабораторная работа №5*	*Вопросы к зачету*
6	Предобработка текстовых данных	LLM-2, MF-4, ML-2	*Лабораторная работа №6*	*Вопросы к зачету*
7	Работа с временными рядами	BD-1, BD-2,	*Лабораторная работа №7*	*Вопросы к зачету*
8	**Комплексная итоговая работа.** Статистический анализ данных (EDA, визуализация, анализ распределений, корреляции, анализ признаков, feature importance). Стратегии по улучшению метрик через данные (feature engineering, отбор признаков, балансировка). Дизайн эксперимента по улучшению данных	LLM-2, MF-4, ML-2	*Лабораторная работа №8*	*Вопросы к зачету*

Показатели, критерии и шкала оценки сформированных компетенций

Соответствие **продвинутому уровню** освоения компетенций планируемым результатам обучения и критериям их оценивания (оценка: **зачтено**):

BD-1 **Способен осуществлять поиск, сбор, очистку и предварительный анализ данных (П)**

BD-1.1 Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи

Знает методы заполнения пропусков в данных и удаления выбросов в табличных данных (случайные величины)
Имеет навыки (**умеет**) очистки зашумленных временных рядов и изображений. Обнаруживает и устраняет выбросы в данных временных рядов. **Владеет** подходами к заполнению пропусков в данных временных рядов и изображений.

BD-1.2 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ
Знает основные методы понижения размерности
Умеет применить основные методы понижения размерности и подбирает оптимальную размерность в зависимости от необходимой доли объяснённой дисперсии.
Владеет методологией применения существующих библиотек, реализующих методы понижения размерности.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных
Знает и умеет применить основы методов отбора признаков и выбирает оптимальное подмножество признаков.
Владеет методологией применения существующих библиотек, реализующих методы отбора признаков.

BD-1.4 **Знает** и умеет применить методы отбора признаков.
Владеет способностью применять методы отбора признаков данных, значимых для исследования.
Умеет отбирать признаки данных, значимые для исследования,
Владеет методами finetuning

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 Знает, как сформировать требования для набора данных. Владеет умениями по формированию требований к наборам и качеству данных для решения задач машинного обучения

BD-2.2 Знает приемы и инструменты для сбора данных из разрозненных источников. Умеет работать с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность. Владеет языками и инструментами для сбора данных и оценки их корректности.

LLM-2 Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения

LLM-2.1 **Понимает принципы fine-tune**
Знает: основные подходы к тонкой настройке: полная настройка всех параметров, поэтапная разморозка слоев, методы эффективной тонкой настройки (P-Tuning, LoRA, QLoRA, Adapter). Гиперпараметры, критически важные для fine-tune: learning rate, scheduler, batch size, и их отличия от обучения с нуля.
Умеет: Отличать дообучение от первичного обучения, знает базовые процедуры **fine-tune,** анализировать задачу и выбирать наиболее подходящий метод fine-tune (полная настройка vs. эффективные методы).
Владеет: **Навыком** осознанного выбора стратегии fine-tune под ограничения

(вычислительные ресурсы, объем данных, требования к качеству). Применяет fine-tune к предобученным моделям на новых датасетах.
• **Методами** анализа и интерпретации процесса дообучения (использование логов, графиков потерь).
• **Критическим мышлением** для оценки целесообразности применения fine-tune в конкретном сценарии versus использования prompt engineering или RAG.

LLM-2.2 **Создаёт обучающие наборы данных.**
Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки.Форматы данных для популярных фреймворков (Hugging Face, TensorFlow, PyTorch) и структур задач (текст-текст, текст-изображение, инстркукции и т.д.). Методы аугментации данных (data augmentation), специфичные для генеративных моделей (e.g., back-translation для текста, модификация промптов). Принципы разбиения данных на обучающую, валидационную и тестовую выборки.
Умеет: Выбирать методы с учетом требований к latency и ресурсам. собирать данные из различных источников: API, веб-скрапинг, открытые датасеты, синтетическая генерация. Очищать и предобрабатывать сырые данные: удаление шума, дубликатов, нормализация текста, приведение к единому формату. Размечать данные в соответствии с поставленной задачей (e.g., составлять пары "инструкция-ответ", аннотировать изображения). Применять методы аугментации данных для увеличения размера и разнообразия обучающего набора.
Владеет: Навыками работы с библиотеками и инструментами для обработки данных (Pandas, NumPy, Hugging Face Datasets).
Методами обеспечения репрезентативности и сбалансированности создаваемого набора данных.
Технологиями создания синтетических данных для задач, где реальных данных недостаточно.
Полным циклом подготовки данных: от сбора сырых данных до формирования готового для обучения объекта (DataLoader, Dataset)

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.
Знает отличия статистического обучения от не статистического, **владеет** классификацией методов статистического машинного обучения. **Умеет** применять и выбирать методы статистического машинного обучения, учитывая особенности данных и задачи, а также объясняет различия между подходами.

MF-4.2 Способен применять статистические методы для построения предсказательных моделей, включая методы для анализа и прогнозирования временных рядов, а также моделирования нестационарных случайных процессов.
Знает: теоретические основы и предположения линейной и логистической регрессии. Понятие стационарности временного ряда, методы проверки (ADF test) и приведения к стационарному виду (дифференцирование, декомпозиция). Классические модели прогнозирования временных рядов (ARIMA, SARIMA, ETS).**Умеет** формализовывать и применять статистические методы идентификации регрессионных и классификационных моделей, понимает

основы базовых вероятностных моделей для временных рядов на основе авторегрессионных зависимостей.. **Владеет** приемами построения модели динамических систем для многомерных временных рядов и полей.

MF-4.3 Способен применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

Знает метрики и меры качества моделей регрессии (в т.ч. на временных рядах), классификации, кластеризации.

Умеет оценивать качество моделей МО

Владеет умением применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

ML-2 Способен применять фундаментальные принципы и методы машинного обучения включая подготовку данных оценку качества моделей и работу с признаками

ML-2.1 Различает основные типы задач машинного обучения и применяет на практике принципы их решения

Знает и различает основные типы задач машинного обучения (обучением с учителем, без учителя и с подкреплением). **Умеет** применить типовые подходы к решению базовых задач с использованием готовых инструментов и библиотек (ScikitLearn) (Б)

Умеет обоснованно применять методы решения задач машинного обучения с учётом характеристик данных и бизнес-контекста, настраивает базовые модели и проводит их оценку (П)

Владеет приемами и инструментами проектирования и реализации комплексных решений машинного обучения для нестандартных задач, включая разработку пайплайнов, оптимизацию моделей и интерпретацию результатов (Э)

Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

4.2.Примеры лабораторных работ и контрольных заданий по разделам учебной дисциплины

Примеры лабораторных работ

Лабораторная работа 1: Планирование и первичны сбор данных для мультимодального датасета

Цель: Научиться формализовать задачу машинного обучения и планировать сбор данных.

Задание:

1. Выберите одну из предложенных задач (или придумайте свою):
   - Прогноз потребления электроэнергии в городе.
   - Определение уровня стресса по данным с носимых устройств.
   - Классификация видов растений по изображениям.

2. Для выбранной задачи:
   - Сформулируйте целевую переменную и тип задачи (регрессия, классификация и т.д.).
   - Составьте список признаков (не менее 10), которые могут быть полезны для решения задачи. Разбейте их по типам (числовые, категориальные, текст, изображения и т.д.).
   - Укажите предполагаемые источники для каждого признака (например, открытые данные, API, краудсорсинг и т.д.).
   - Оцените объем данных, необходимых для обучения модели, и предложите стратегию разметки данных.
3. Напишите отчет, включающий:
   - Описание задачи.
   - Обоснование выбранных признаков.
   - План сбора данных.
   - Оценку трудозатрат и рисков.

Критерии оценки:

- Полнота и адекватность плана (0-5 баллов);
- Обоснованность выбора признаков (0-5 баллов);
- Реалистичность плана сбора и разметки (0-5 баллов).

Лабораторная работа 2: Сбор и первичная оценка мультимодальных данных - первичная обработка и очистка данных

Цель: Освоить основные техники обработки и очистки данных.

Задание:

1. Вам предоставлен датасет (например, с Kaggle) с различными типами признаков и проблемами (пропуски, выбросы, дубликаты и т.д.).
2. Проведите первичную очистку данных:
   - Удалите или обработайте дубликаты.
   - Обработайте пропущенные значения (удаление, интерполяция, заполнение средним/медианой и т.д.) с обоснованием выбора метода.
   - Найдите и обработайте выбросы (методами IQR, Z-score и т.д.).
3. Проведите первичный анализ данных:
   - Постройте гистограммы распределений числовых признаков.
   - Постройте боксплоты для обнаружения выбросов.
   - Создайте таблицу с основными статистиками (среднее, медиана, стандартное отклонение и т.д.).
4. Напишите отчет с описанием проведенных действий и обоснованием выбранных методов.

Критерии оценки:

1. Корректность обработки пропусков и выбросов (0-5 баллов)
2. Глубина анализа (0-5 баллов)
3. Качество визуализаций (0-5 баллов)

Лабораторная работа 3: Предобработка числовых и категориальных данных

Цель: Научиться проводить предобработку числовых и категориальных данных.

Задание:

1. Работа с числовыми данными:
   Проведите масштабирование и нормализацию числовых признаков (MinMaxScaler, StandardScaler, RobustScaler).
   Создайте новые признаки (полиномиальные, бининг, логарифмирование и т.д.).
2. Работа с категориальными данными:
   Закодируйте категориальные признаки (One-Hot Encoding, Label Encoding, Target Encoding).
   Объедините редкие категории.
3. Создайте пайплайн предобработки данных с использованием ColumnTransformer и Pipeline из sklearn.
4. Обучите простую модель (например, логистическую регрессию или дерево решений) на обработанных данных и сравните ее производительность с моделью, обученной на необработанных данных.

Критерии оценки:

1. Разнообразие и корректность методов предобработки (0-5 баллов)
2. Качество пайплайна (0-5 баллов)
3. Сравнение производительности моделей (0-5 баллов)

Лабораторная работа 4: Пайплайн аугментации изображений

Цель: Научиться проводить аугментацию и предобработку изображений.

Задание:

1. Возьмите небольшой датасет изображений (например, CIFAR-10 или набор с изображениями кошек и собак).
2. Реализуйте пайплайн предобработки изображений:
   - Изменение размера.
   - Нормализация пикселей.
   - Перевод в grayscale (если необходимо).
3. Реализуйте аугментацию данных:
   - Повороты, сдвиги, отражения.
   - Изменение яркости, контраста.
   - Добавление шума.

   Реализовать 8+ различных аугментаций, визуализировать результаты аугментации
4. Создать сбалансированный датасет
5. Обучите простую сверточную нейронную сеть (CNN) на исходных и аугментированных данных и сравните результаты.

Критерии оценки:

1. Разнообразие аугментаций (0-2)
2. Качество предобработки (0-2 баллов)
3. Сравнение производительности моделей (0-3 баллов)
4. Балансировка датасета (0-3)

Лабораторная работа 5: Предобработка аудиоданных, извлечение аудиопризнаков

Цель: Освоить основные методы предобработки аудиоданных.

Задание:

1. Возьмите аудиодатасет (например, набор звуковых команд или музыкальных жанров).
2. Проведите предобработку аудио:
   - Ресемплирование.
   - Нормализация громкости.
   - Обрезка тишины.

3. Извлеките признаки из аудио:
   - MFCC.
   - Spectrogram.
   - Chroma features.

   Извлечь 5+ типов аудио признаков, реализовать 3+ метода аугментации
   Визуализировать спектрограммы и MFCC, создать датасет фиксированного размера
4. Обучите модель (например, SVM или простую нейронную сеть) на извлеченных признаках и оцените качество.

Критерии оценки:

1 Полнота извлечения признаков (0-3)
2 Качество аугментации (0-3)
3 Визуализация результатов (0-2)
4 Организация датасета (0-2)

Лабораторная работа 6: Пайплайн обработки текста

Цель: Освоить основные методы предобработки текстовых данных.

Задание:

1. Возьмите текстовый датасет (например, отзывы с IMDb или твиты).
2. Проведите предобработку текстов:
   1. Очистка (удаление HTML-тегов, пунктуации, чисел и т.д.).
   2. Токенизация.
   3. Удаление стоп-слов.
   4. Лемматизация или стемминг.
3. Векторизуйте тексты с помощью:
   1. Bag of Words (CountVectorizer).
   2. TF-IDF (TfidfVectorizer).
   3. Word2Vec или GloVe (предобученные модели).
4. Обучите модель классификации (например, Naive Bayes или логистическую регрессию) на полученных векторах и сравните результаты между разными методами векторизации.

Критерии оценки:

1. Качество предобработки текста (0-5 баллов)
2. Корректность векторизации (0-5 баллов)
3. Сравнение методов векторизации (0-5 баллов)

Лабораторная работа 7: Работа с временными рядами

Цель: Научиться обрабатывать временные ряды и извлекать из них признаки.

Задание:

1. Возьмите датасет с временными рядами (например, продажи магазина или курс акций).
2. Проведите предобработку:
   1. Заполнение пропусков (интерполяция, заполнение предыдущим значением).
   2. Сглаживание (скользящее среднее, экспоненциальное сглаживание).
   3. Проверка на стационарность (тест Дики-Фуллера) и приведение к стационарности (дифференцирование).
3. Извлеките признаки:
   1. Лаговые признаки.
   2. Статистики по окнам (среднее, стандартное отклонение и т.д.).
   3. Дата-признаки (день недели, месяц, праздники и т.д.).
4. Обучите модель (например, ARIMA, LSTM или линейную регрессию) на обработанных данных.

Критерии оценки:

1. Корректность предобработки (0-5 баллов)

2. Разнообразие извлеченных признаков (0-5 баллов)
3. Качество модели (0-5 баллов)

Лабораторная работа 8: Итоговая комплексная работа. Настройка и отбор признаков. Балансировка данных и стратегии семплирования

Цель: Научиться создавать новые признаки и отбирать наиболее информативные. Научиться работать с несбалансированными данными.

Задание:

1.На основе датасета из предыдущих работ (или нового) создайте новые признаки:

1. На основе доменных знаний.
2. Путем взаимодействия существующих признаков.
3. Используя преобразования (полиномиальные, логарифмические и т.д.).

2. Проведите отбор признаков:

1. Методы фильтров (корреляция, взаимная информация).
2. Методы обертки (Recursive Feature Elimination).
3. Методы встроенные (Lasso, важность признаков в деревьях).

3. Сравните производительность модели на исходном наборе признаков и на наборе после отбора.

1. На основе несбаллансированного датасет примените методы балансировки:
2. Случайное недосэмплирование и пересэмплирование.
3. SMOTE и его варианты.
4. ADASYN.

4. Обучите модель на исходных данных и на сбалансированных и сравните метрики (Precision, Recall, F1-score, ROC-AUC).

В ходе лабораторных работ студентам необходимо использовать платформу MLflow для отслеживания экспериментов по предобработке данных и сравнении влияния разных методов на итоговую модель.

Предлагаемые направления использования:

1. **Отслеживание этапов предобработки данных**:
   - Студенты могут использовать MLflow для записи параметров предобработки (например, стратегия заполнения пропусков, метод кодирования категориальных переменных, стандартизация),
   - Логировать артефакты: обработанные датасеты, графики распределений, код преобразований.
2. **Сравнение различных методов предобработки**:
   - Запускать несколько экспериментов с разными методами предобработки и логировать метрики модели (например, accuracy, F1-score) после обучения на предобработанных данных.
   - Через UI MLflow сравнивать, как разные методы влияют на результат.
3. **Воспроизводимость**:
   - MLflow Projects позволяет зафиксировать код и окружение, чтобы любой студент мог воспроизвести эксперимент.
4. **Управление моделями**:
   - Сохранять модели, обученные на данных, обработанных разными способами, и регистрировать их в MLflow Model Registry.
   - Сравнивать модели и переводить лучшие в продакшн (условно, для демонстрации).

5. **Примерный план лабораторной работы**:
   - Разработать несколько сценариев предобработки данных.
   - Для каждого сценария:
     a) Запустить эксперимент в MLflow.
     b) Выполнить предобработку с определенными параметрами.
     c) Обучить простую модель (например, логистическая регрессия или случайный лес).
     d) Записать параметры предобработки, метрики модели и сохранить обработанный датасет как артефакт.
   - Сравнить эксперименты в UI и выбрать лучший метод.
6. **Интеграция с инструментами**:
   - Показать, как MLflow интегрируется с библиотеками для работы с данными (Pandas, Scikit-learn) и визуализации (Matplotlib, Seaborn).
7. **Демонстрация полного цикла**:
   - От сырых данных до модели, с отслеживанием каждого шага.

Критерии оценки лабораторных работ:

1. Количество и качество новых признаков (0-2 баллов)
2. Обоснованность отбора признаков (0-2 баллов)
3. Корректность применения методов балансировки (0-3 баллов)
4. Сравнение метрик и производительности (0-3 баллов)

Критерии оценивания лабораторных работ:

«неудовлетворительно» – 1–2 балла – испытывает трудности применения теоретических знаний к решению практических задач; допускает принципиальные ошибки в выполнении заданий;

«удовлетворительно» – 2–3 баллов – применяет теоретические знания к решению заданий в контрольной задаче; справляется с выполнением типовых практических задач по известным алгоритмам, правилам, методам;

«хорошо» – 4 балла – правильно применяет теоретические знания к решению заданий в контрольной задаче; выполняет типовые практические задания на основе адекватных методов, способов, приемов, решает задания повышенной сложности, допускает незначительные отклонения;

«отлично» – 5 баллов – творчески применяет знания теории к решению заданий в контрольной задаче, находит оптимальные решения для выполнения практического задания; свободно выполняет типовые практические задания на основе адекватных методов, способов, приемов; решает задания повышенной сложности, находит нестандартные решения в проблемных ситуациях.

Перечень компетенций, проверяемых оценочным средством: BD-1, BD-2, LLM-2, ML-2, MF-4.

4.3. Примеры контрольных заданий для промежуточной аттестации и на зачет

Раздел 1: Планирование датасета под задачу
Задание 1.1 (Базовое)
Спроектируйте датасет для системы рекомендации фильмов на основе предпочтений пользователя. Составьте таблицу с:

- 10+ гипотетическими признаками разных типов;
- Источниками получения каждого признака;
- Оценкой стоимости сбора.

Задание 1.2 (Продвинутое)
Для задачи предсказания оттока клиентов банка:

- Составьте матрицу "риск оттока - доступность данных"
- Предложите 3 альтернативных схемы датасета с разным бюджетом
- Обоснуйте, какие признаки будут иметь наибольшую предсказательную силу.

Раздел 2: Первичный сбор и оценка качества
Задание 2.1 (Практическое)
Дан набор из 1000 JSON-файлов с данными о пользователях. Напишите скрипт, который:

- Проверяет целостность всех файлов;
- Валидирует обязательные поля в каждом файле;
- Создает сводный отчет о качестве данных.

Задание 2.2 (Аналитическое)
Проанализируйте предоставленный датасет с изображениями товаров. Составьте отчет о:

- Процент бракованных изображений (битые, слишком темные/светлые);
- Распределение по разрешениям и форматам;
- Рекомендации по улучшению качества сбора.

Раздел 3: Предобработка табличных данных
Задание 3.1 (Кодирование)
Дан датасет с категориальными признаками: "цвет" (10 значений), "размер" (порядковый), "город" (150 значений). Предложите и обоснуйте:

- Схему кодирования для каждого признака
- Параметры для OneHotEncoder, OrdinalEncoder, TargetEncoder
- Спроектируйте пайплайн предобработки

Задание 3.2 (Обработка выбросов)
Реализуйте 3 различных метода обработки выбросов для числовых признаков:

- IQR метод
- Z-score
- Isolation Forest

Сравните их влияние на метрики линейной регрессии.

Раздел 4: Предобработка изображений
Задание 4.1 (Аугментация)
Создайте пайплайн аугментации для датасета медицинских снимков (рентген), который:

- Увеличивает датасет в 4 раза;
- Сохраняет медицинскую значимость изображений;
- Включает как geometric, так и photometric трансформации.

Задание 4.2 (Transfer Learning)
Подготовьте датасет изображений собак и кошек для Transfer Learning на ResNet50:

- Приведите все изображения к нужному размеру;
- Примените нормализацию ImageNet;

- Создайте генератор данных с аугментацией.

Раздел 5: Предобработка аудио

Задание 5.1 (Feature Extraction)

Напишите функцию, которая извлекает из аудиофайла:

- MFCC (13, 26, 39 коэффициентов);
- Mel-спектрограмму;
- Chroma features;
- Spectral contrast.

Задание 5.2 (Аугментация)

Реализуйте 3 метода аугментации для аудио датасета распознавания команд:

- Time stretching;
- Pitch shifting;
- Добавление шума.

Оцените их влияние на точность модели.

Раздел 6: Предобработка текста

Задание 6.1 (Токенизация)

Сравните 3 подхода к токенизации для датасета новостей:

- Word-level с стоп-словами;
- Subword (BPE);
- С использованием предобученного BERT токенизатора.

Задание 6.2 (Векторизация)

Для датасета отзывов на товары создайте и сравните:

- TF-IDF векторизацию;
- Word2Vec эмбеддинги;
- BERT эмбеддинги.

Проанализируйте их эффективность для задачи классификации тональности.

Раздел 7: Временные ряды

Задание 7.1 (Сегментация временных рядов)

Сравните 3 подхода к сегментации временных рядов для датасета с акселерометра:

- Фиксированное окно с перекрытием;
- Сегментация на основе важных точек (пики, впадины);
- Сегментация с помощью предобученной модели (например, CNN).

Задание 7.2 (Векторизация временных рядов)

Для датасета временных рядов с датчиков вибрации создайте и сравните:

- Признаки, извлеченные с помощью вейвлет-преобразования;
- Признаки, извлеченные с помощью автоэнкодера;
- Признаки, извлеченные с помощью предобученной модели (например, сверточной сети, обученной на ImageNet для спектрограмм).

Проанализируйте их эффективность для задачи классификации неисправностей оборудования.

Раздел 8: Статистический анализ и Feature Engineering

Задание 8.1 (Стат-анализ)

Проведите полный статистический анализ датасета с таймсериями датчиков:

- Проверьте нормальность распределений;
- Проанализируйте автокорреляцию;
- Найдите статистически значимые различия между классами.

Задание 8.2 (Feature Engineering)
Создайте 10 новых признаков для датасета транзакций:

- 5 на основе доменных знаний;
- 3 с использованием статистических методов;
- 2 с помощью анализа временных рядов;

Оцените их важность с помощью Random Forest.

Задание 8.3 (A/B тестирование)
Спроектируйте эксперимент по сравнению 3 версий датасета:

- Исходная версия;
- С применением SMOTE;
- С дополнительными синтетическими данными.

Определите размер выборки и критерии статистической значимости.
Задание 8.4 (Оптимизация)
Для фиксированной модели LogisticRegression предложите и проверьте 5 стратегий улучшения данных:

- Балансировка классов;
- Подбор аугментации;
- Feature selection;
- Нормализация признаков;
- Создание полиномиальных признаков.

Комплексные контрольные работы, выделяемое на решение время: от *120 до 180 минут*

Контрольная работа 1 (Средний уровень)
Задача: Подготовьте данные для системы определения спама в SMS
Требования:

1. Загрузите и проведите первичный анализ датасета;
2. Реализуйте пайплайн очистки текста;
3. Сравните 2 метода векторизации;
4. Проведите статистический анализ признаков;
5. Улучшите метрики на 5% только через работу с данными.

Контрольная работа 2 (Продвинутый уровень)
Задача: Мультимодальная классификация эмоций по видео
Требования:

1. Спроектируйте схему датасета (видео + аудио + текст)
2. Реализуйте пайплайны для каждого типа данных
3. Проведите анализ качества и согласованности разметки
4. Создайте комбинированные признаки
5. Добейтесь улучшения метрик на 10% через усовершенствование датасета

Критерии оценки контрольных работ

Для практических заданий (0-10 баллов):

- **0-3 балла:** Решение содержит критические ошибки, не работает
- **4-6 баллов:** Решение работает, но не оптимально, нет анализа
- **7-8 баллов:** Решение корректное, есть базовый анализ
- **9-10 баллов:** Решение оптимальное, глубокий анализ, обоснование выбора методов

Для теоретических заданий (0-10 баллов):

- **0-3 балла:** Поверхностный ответ, без примеров и обоснования
- **4-6 баллов:** Основные понятия раскрыты, но нет глубины
- **7-8 баллов:** Полный ответ с примерами, но без критического анализа
- **9-10 баллов:** Глубокий ответ с анализом преимуществ/недостатков методов

Шкала оценок:

- **0-49 баллов:** Неудовлетворительно
- **50-69 баллов:** Удовлетворительно
- **70-89 баллов:** Хорошо
- **90-100 баллов:** Отлично

Пример решения задания 3.1

python

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder, OrdinalEncoder
from category_encoders import TargetEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Обоснование выбора методов кодирования:
"""
Цвет (10 значений) -> OneHot Encoding:
Номинальный признак, малое количество уникальных значений, не создает слишком много новых признаков

Размер (порядковый) -> Ordinal Encoding:
Явный порядок значений, сохраняет информацию о рангах

Город (150 значений) -> Target Encoding:
Большое количество категорий, OneHot создал бы 150 новых признаков, есть риск переобучения, нужна регуляризация:

preprocessor = ColumnTransformer([
   ('color', OneHotEncoder(drop='first', sparse=False), ['color']),
   ('size', OrdinalEncoder(), ['size']),
   ('city', TargetEncoder(smoothing=10), ['city'])
], remainder='passthrough')

pipeline = Pipeline([
   ('preprocessor', preprocessor),
   ('scaler', StandardScaler()),
   ('model', LogisticRegression())
])
```

Задачи для сбора и предобработки данных

Задача 1: "Сбор и обработка данных из API социальной сети"

Описание задачи:

Необходимо собрать данные о пользователях из API социальной сети и подготовить их для анализа поведения пользователей.

Сбор данных:

Напишите скрипт для получения данных пользователей через API

Соберите информацию о 1000 пользователей (id, имя, дата регистрации, количество друзей, количество постов). Соберите дополнительные данные о активностях пользователей за последние 30 дней.

Обработка сырых данных:

Объедините данные из разных API endpoints в единую таблицу. Обработайте вложенные JSON структуры (списки друзей, истории активностей). Сохраните собранные данные в SQLite базу данных

Предобработка:

Нормализуйте имена пользователей (приведите к единому регистру, удалите лишние пробелы).Рассчитайте производные признаки:

activity_rate (посты в день)

registration_period (дней с момента регистрации)

friends_growth_rate (скорость роста числа друзей)

Работа с пропусками:

Разработайте стратегию обработки пропущенных данных о активностях

Используйте интерполяцию для заполнения пропусков в временных рядах

Создайте флаги для обозначения импутированных значений

Задача 2 (для проектной деятельсти): "Подготовка текстовых данных из внешних источников"

Описание задачи:

Подготовьте данные для модели прогнозирования цен на недвижимость, объединив данные из разных источников.

Источники данных: Открытые источники с объявлениями о продаже квартир

API карт для получения геоданных, **в**нешние CSV файлы с инфраструктурой районов

Текстовые описания из объявлений.

Задания:

Сбор дополнительных данных:

Напишите скрипт для получения расстояний до метро через Geocoding API.Соберите данные о количестве парков, школ, магазинов в радиусе 1 км. Извлеките данные о ремонте из текстовых описаний с помощью регулярных выражений

Объединение данных:

Создайте единый датафрейм, объединив данные из 2-3 источников. Реализуйте различные типы JOIN'ов для сохранения максимального количества наблюдений. Обработайте случаи расхождения в ключах объединения

Обработка текстовых данных: Примените очистку текста (удаление стоп-слов, лемматизация).

Извлеките признаки из описаний:

Наличие ремонта (капитальный, косметический, без ремонта)

Упоминания особенностей (вид из окна, мебель, техника)

Создайте мешок слов для ключевых характеристик

Создание сложных признаков:

Рассчитайте price_per_sqm (цена за кв. метр)

Создайте признак infrastructure_score на основе собранных геоданных

Постройте признаки, описывающие соотношение комнат и площади
Создайте временные признаки (сезонность продаж)

Валидация данных:

Напишите тесты для проверки качества данных:

Проверка на дубликаты
Валидация диапазонов значений
Проверка согласованности связанных полей
Создайте пайплайн предобработки, который можно применять к новым данным

Перечень компетенций, проверяемых оценочным средством: BD-1, BD-2, LLM-2, ML-2, MF-4.

Зачетно-экзаменационные материалы для промежуточной аттестации (зачет)

1. Этапы планирования датасета под задачу машинного обучения. Критерии оценки качества спроектированного датасета.
2. Методы сбора данных из различных источников (API, веб-скрапинг, базы данных). Преимущества и недостатки каждого.
3. Оценка качества данных: основные метрики и методы для разных типов данных (таблицы, изображения, аудио, текст).
4. Предобработка табличных данных: обработка пропущенных значений, выбросов, кодирование категориальных признаков.
5. Методы масштабирования и нормализации данных. В каких случаях применяется каждый метод?
6. Предобработка изображений: аугментация данных, нормализация, изменение размера. Методы сохранения разметки при аугментации.
7. Предобработка аудиоданных: основные этапы (ресемплирование, обрезка тишины, извлечение признаков). Популярные признаки для аудио (MFCC, спектрограммы и др.).
8. Предобработка текстовых данных: токенизация, очистка текста, векторизация (TF-IDF, Word2Vec, BERT). Сравнение методов.
9. Статистический анализ данных: проверка нормальности, анализ корреляций, статистические тесты для отбора признаков.
10. Feature Engineering: создание новых признаков для табличных данных, временных рядов, изображений, аудио и текста.
11. Методы борьбы с дисбалансом классов (SMOTE, взвешенные функции потерь, ансамбли). Преимущества и недостатки.
12. Стратегии разделения данных на обучение, валидацию и тест. Кросс-валидация и ее виды.
13. Дизайн экспериментов по улучшению данных: A/B тестирование, ablation study, оценка статистической значимости.
14. Пайплайны предобработки данных: преимущества использования, реализация в scikit-learn, обработка разных типов признаков.
15. Оценка качества предобработки данных: метрики качества модели до и после улучшения данных, анализ ошибок.

4.4 Методические рекомендации к сдаче зачета и критерии оценки ответа

Промежуточная аттестация традиционно служат основным средством обеспечения в учебном процессе «обратной связи» между преподавателем и обучающимся, необходимой для стимулирования работы обучающихся и совершенствования методики преподавания учебных

дисциплин. Итоговой формой контроля сформированности компетенций, обучающихся по дисциплине «Математические модели нейронных сетей» является зачет. Студенты обязаны сдать зачет в соответствии с расписанием и учебным планом. Зачет по дисциплине преследует цель оценить работу студента за курс, получение теоретических знаний, их прочность, развитие творческого мышления, приобретение навыков самостоятельной работы, умение применять полученные знания для решения практических задач и является формой контроля усвоения студентом учебной программы по дисциплине, выполнения практических, контрольных, реферативных работ. Форма проведения зачета: устно. Результат сдачи зачета по прослушанному курсу должен оцениваться как итог деятельности студента в семестре, а именно – по посещаемости лекций, результатам работы на лекционных и практических занятиях, прохождения тестовых заданий, решения расчетно-графических заданий и задач, выполнения контролируемой самостоятельной работы. Студенты, прошедшие все виды испытаний, предусмотренных оценочными средствами положительно (т.е. по каждому виду оценочных средств были получены оценки «удовлетворительно», и(или) «хорошо», и(или) «отлично») выставляется «зачтено». При этом допускается на очной форме обучения пропуск не более 20% занятий, с обязательной отработкой пропущенных семинаров. Студенты, у которых количество пропусков, превышает установленную норму, не выполнившие все виды работ и неудовлетворительно работавшие в течение семестра, проходят собеседование с преподавателем, в виде устного ответа на один теоретический вопрос и решения одного расчетно-графического задания. Преподавателю предоставляется право задавать студентам дополнительные вопросы по всей учебной программе дисциплины. Результат сдачи зачета заносится преподавателем в ведомость и зачетную книжку.

Критерии оценки зачета. Оценка «зачтено» выставляется студенту, если дан полный развёрнутый ответ на теоретический вопрос, логически правильно изложены ответы на дополнительные вопросы; студент показал умение свободно выполнять расчетно-графическое задание, предусмотренное дисциплиной, самостоятельность решения задания и приводимых суждений; все расчеты сделаны правильно; выводы вытекают из содержания задания, предложения обоснованы, в изложении ответов нет существенных недостатков. В то же время в ответе могут присутствовать незначительные фактические ошибки в изложении материала. Оценка «не зачтено» выставляется при несоответствии ответа заданному вопросу, наличии грубых ошибок, использовании при ответе ненадлежащих источников; студент показал пробелы в знаниях основного учебного материала, значительные пробелы в знаниях теоретических компонентов программы; неумение ориентироваться в основных научных теориях и концепциях, связанных с осваиваемой дисциплиной, неточное их описание; слабое владение научной терминологией и профессиональным инструментарием; допустил принципиальные ошибки в выполнении предусмотренной дисциплиной практического задания, изложение ответа на вопросы с существенными лингвистическими и логическими ошибками.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:
– в печатной форме увеличенным шрифтом,
– в форме электронного документа.
Для лиц с нарушениями слуха:
– в печатной форме,
– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:
– в печатной форме,
– в форме электронного документа.
Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

4.5. Методические указания по организации вычислительной инфраструктуры

Требования к аппаратному и программному обеспечению рабочих мест

Аппаратные требования

Для выполнения лабораторных работ по изучению методов подготовки данных машинного обучения студентам и преподавателю необходим стационарный компьютер или ноутбук с современной конфигурацией. Рекомендуется многопроцессорный CPU, например Intel Core i3/i5/i7 не ниже 4-го поколения или аналогичный AMD Ryzen, с поддержкой многопоточности и оперативной памятью не менее 8 ГБ. Для работы с GPU-вычислениями требуется видеокарта NVIDIA для CUDA или совместимая с OpenCL/ROCm. Компьютер должен иметь стабильное подключение к сети Интернет со скоростью не ниже 5–10 Мбит/с для скачивания SDK, библиотек и обновлений программного обеспечения.

Программные требования.

На рабочих станциях должна быть установлена современная операционная система, включая Windows 10 или 11, актуальные версии macOS или дистрибутивы GNU/Linux, при этом системы должны регулярно обновляться для поддержания безопасности и совместимости с инструментами курса. Для разработки необходимы библиотеки, указанные в разделе «Инструменты и библиотеки».

Студенты также должны иметь доступ к системе контроля версий Git, интерпретатору Python версии 3.10 и выше с менеджером пакетов pip или conda для анализа результатов и построения графиков, при необходимости с установкой Jupyter Notebook/Lab. Для локального тестирования и отладки программ может использоваться Docker, при этом на Windows требуется Docker Desktop с WSL2, а на Linux и macOS платформа поддерживается напрямую. Все программное обеспечение должно быть настроено так, чтобы студенты имели доступ ко всем инструментам во время лабораторных работ, а преподаватель мог управлять инфраструктурой и контролировать результаты, включая репозитории, CI/CD и тестирование. Необходимо обеспечить разрешение исходящих подключений по HTTPS, открытые порты 80 и 443, а также наличие прав на установку программного обеспечения или взаимодействие с системным администратором для их установки.

Инструменты и библиотеки:

Категория	Python	R
Основные ML	scikit-learn	caret, tidymodels
Обработка данных	pandas, numpy	dplyr, data.table
Визуализация	matplotlib, seaborn, plotly	ggplot2, plotly

Категория	Python	R
Ансамбли	xgboost, lightgbm, catboost	randomForest, xgboost, gbm
Бустинг	xgboost, lightgbm, catboost	xgboost, gbm
Деревья решений	scikit-learn	rpart
Нейросети	tensorflow, keras, pytorch	nnet, keras
SVM	scikit-learn	e1071
Линейные модели	scikit-learn, statsmodels	glm, lm
Интерпретация моделей	shap, eli5, lime	DALEX, iml
Оптимизация гиперпараметров	optuna, scikit-optimize	mlrMBO, tune
Несбалансированные данные	imbalanced-learn	ROSE, smotefamily
Работа с текстом	nltk, spaCy	tm, tidytext
Верификация моделей	scikit-learn	ROCR, mlbench
Пайплайны	scikit-learn	recipes
Даты и время	pandas	lubridate
Строки	pandas	stringr
Эксперименты	mlflow	MLflow

Исходные данные: готовые датасеты, данные собранные в ходе выполнения работ.

5. Перечень основной и дополнительной учебной литературы, информационных ресурсов и технологий необходимых для освоения дисциплины

5.1 Основная литература

(в том числе публикации конференций A*)

1. Митяков, Е. С. Искусственный интеллект и машинное обучение : учебное пособие для вузов / Е. С. Митяков, А. Г. Шмелева, А. И. Ладынин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — 252 с. — ISBN 978-5-507-51198-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/507451 (дата обращения: 24.10.2025). — Режим доступа: для авториз. пользователей.
2. Баланов, А. Н. Машинное обучение и искусственный интеллект : учебное пособие для вузов / А. Н. Баланов. — 2-е изд., стер. — Санкт-Петербург : Лань, 2025. — 172 с. — ISBN 978-5-507-52891-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/462248 (дата обращения: 24.10.2025). — Режим доступа: для авториз. пользователей.
3. Машинное обучение : учебник : [16+] / Е. Ю. Бутырский, В. В. Цехановский, Н. А. Жукова [и др.]. – Москва : Директ-Медиа, 2023. – 368 с. : ил., табл., схем., граф. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=701807 (дата обращения: 24.10.2025). – Библиогр. в кн. – ISBN 978-5-4499-3778-0. – DOI 10.23681/701807. – Текст : электронный.
4. Биомедицинские сигналы и изображения в цифровом здравоохранении : хранение, обработка и анализ : учебное пособие / А. П. Немирко, Л. А. Манило, А. Ю. Долганов [и др.] ; под общ. ред. В. С. Кубланова ; Уральский федеральный университет им. первого Президента России Б. Н. Ельцина. – Екатеринбург : Издательство Уральского университета, 2020. – 243 с. : схем., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=698902 (дата обращения: 24.10.2025). – Библиогр. в кн. – ISBN 978-5-7996-2990-8. – Текст : электронный.

2. Целых, А. Н. Применение временных рядов для анализа больших данных : учебное пособие по курсу «Математические методы анализа больших данных» : [16+] / А. Н. Целых, В. С. Васильев, Э. М. Котов ; Южный федеральный университет. – Ростов-на-Дону ; Таганрог : Южный федеральный университет, 2021. – 86 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=691448 (дата обращения: 24.10.2025). – Библиогр. в кн. – ISBN 978-5-9275-3983-3. – Текст : электронный.
3. Целых, А. Н. Современные методы прикладной информатики в задачах анализа данных : учебное пособие по курсу «Методы интеллектуального анализа данных» : [16+] / А. Н. Целых, А. А. Целых, Э. М. Котов ; Южный федеральный университет. – Ростов-на-Дону ; Таганрог : Южный федеральный университет, 2021. – 130 с. : ил., табл., схем. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=683920 (дата обращения: 24.10.2025). – Библиогр. в кн. – ISBN 978-5-9275-3783-9. – Текст : электронный.
4. Sun, X., Li, J., Kovalenko, A.V., Feng, W., Ou, Y. Integrating Reinforcement Learning and Learning From Demonstrations to Learn Nonprehensile Manipulation //IEEE Transactions on Automation Science and Engineering, 2023, 20(3), 1735–1744, DOI: 10.1109/TASE.2022.3185071, Q1
5. Petukhova, A.V.; Kovalenko, A.V.; Ovsyannikova, A.V. Algorithm for Optimization of Inverse Problem Modeling in Fuzzy Cognitive Maps. Mathematics 2022, 10, 3452. DOI: 10.3390/math10193452, Q1
6. Kadurin, Artur, et al. "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology." Oncotarget 8.7 (2016): 10883.
7. Kadurin, Artur, et al. "druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico." Molecular pharmaceutics 14.9 (2017): 3098-3104.
8. Polykovskiy, Daniil, et al. "Molecular sets (MOSES): a benchmarking platform for molecular generation models." Frontiers in pharmacology 11 (2020): 565644.
9. Khrabrov, Kuzma, et al. "$\nabla^ 2$ DFT: A Universal Quantum Chemistry Dataset of Drug-Like Molecules and a Benchmark for Neural Network Potentials." Advances in Neural Information Processing Systems 37 (2024): 36869-36889.
10. Haoran Xu, Philipp Koehn, and Kenton Murray. 2022. The Importance of Being Parameters: An Intra-Distillation Method for Serious Gains. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 170–183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
11. Wai Ching Leung, Shira Wein, and Nathan Schneider. 2022. Semantic Similarity as a Window into Vector- and Graph-Based Metrics. In Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 106–115, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
12. Anna Lorincz, David Graus, Dor Lavi, and Joao Lebre Magalhaes Pereira. 2022. Transfer learning for multilingual vacancy text generation. In Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 207–222, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics

5.2. Дополнительная литература:

1. Разметка данных в машинном обучении: процесс, разновидности и рекомендации [Электронный ресурс]. - URL: https://habr.com/ru/articles/678524/. - (Дата обращения: 10.10.2025).

2. Неструктурированные данные: примеры, инструменты, методики и рекомендации [Электронный ресурс]. - URL: https://habr.com/ru/articles/756454/. -(Дата обращения: 10.10.2025).
3. Structured vs. Unstructured Data: What's the Difference? [Электронный ресурс]. - URL: https://www.coursera.org/articles/structured-vs-unstructured-data. - (Дата обращения: 10.10.2025).
4. What is unstructured data? [Электронный ресурс]. - URL: https://www.elastic.co/what-is/unstructured-data. - (Дата обращения: 10.10.2025).
5. Kovriguina, L., Shilin, I., Putintseva, A., Shipilo, A. Multilevel Annotation in the Corpus for Parsing Russian Spontaneous Speech. In: Karpov, A., Jokisch, O., Potapova, R. (eds) Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science(), vol 11096. Springer, 2018 - 311-320 p.
6. Anthony S. Training Data for Machine Learning. O'Reilly Media, 2023. - 332 p. books on Data Annotation [Электронный ресурс]. - URL: https://www.aistartups.org/books/data_annotation/. - (Дата обращения: 01.10.2025).
7. Захаров В., Богданова С. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

5.3. Интернет-ресурсы, в том числе современные профессиональные базы данных, информационные справочные системы и конференции

Конференции A*:

1. https://openreview.net/forum?id=FMMF1a9ifL
2. https://openreview.net/forum?id=ElUrNM9U8c#discussion
3. https://openreview.net/forum?id=JoO6mtCLHD
4. https://aclanthology.org/2024.findings-emnlp.760/
5. https://aclanthology.org/2020.coling-main.588/
6. https://link.springer.com/chapter/10.1007/978-3-030-72113-8_30
7. https://link.springer.com/chapter/10.1007/978-3-031-42448-9_10
8. https://aclanthology.org/2024.findings-naacl.288/

Электронно-библиотечные системы (ЭБС):

1. ЭБС «ЮРАЙТ» https://urait.ru/
2. ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН» http://www.biblioclub.ru/
3. ЭБС «BOOK.ru» https://www.book.ru
4. ЭБС «ZNANIUM.COM» www.znanium.com
5. ЭБС «ЛАНЬ» https://e.lanbook.com

Профессиональные базы данных

1. Scopus http://www.scopus.com/
2. ScienceDirect https://www.sciencedirect.com/
3. Журналы издательства Wiley https://onlinelibrary.wiley.com/
4. Научная электронная библиотека (НЭБ) http://www.elibrary.ru/
5. Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН http://archive.neicon.ru
6. Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/
7. Президентская библиотека им. Б.Н. Ельцина https://www.prlib.ru/

8. База данных CSD Кембриджского центра кристаллографических данных (CCDC) https://www.ccdc.cam.ac.uk/structures/
9. Springer Journals: https://link.springer.com/
10. Springer Journals Archive: https://link.springer.com/
11. Nature Journals: https://www.nature.com/
12. Springer Nature Protocols and Methods: https://experiments.springernature.com/sources/springer-protocols
13. Springer Materials: http://materials.springer.com/
14. Nano Database: https://nano.nature.com/
15. Springer eBooks (i.e. 2020 eBook collections): https://link.springer.com/
16. "Лекториум ТВ" http://www.lektorium.tv/
17. Университетская информационная система РОССИЯ http://uisrussia.msu.ru

Информационные справочные системы

1. **Консультант Плюс** - справочная правовая система (доступ по локальной сети с компьютеров библиотеки)

Ресурсы свободного доступа

1. КиберЛенинка http://cyberleninka.ru/;
2. Американская патентная база данных http://www.uspto.gov/patft/
3. Министерство науки и высшего образования Российской Федерации https://www.minobrnauki.gov.ru/;
4. Федеральный портал "Российское образование" http://www.edu.ru/;
5. Информационная система "Единое окно доступа к образовательным ресурсам" http://window.edu.ru/;
6. Единая коллекция цифровых образовательных ресурсов http://school-collection.edu.ru/;
7. Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском" https://pushkininstitute.ru/;
8. Справочно-информационный портал "Русский язык" http://gramota.ru/;
9. Служба тематических толковых словарей http://www.glossary.ru/;
10. Словари и энциклопедии http://dic.academic.ru/;
11. Образовательный портал "Учеба" http://www.ucheba.com/;
12. Законопроект "Об образовании в Российской Федерации". Вопросы и ответы http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety.

Собственные электронные образовательные и информационные ресурсы КубГУ

1. Электронный каталог Научной библиотеки КубГУ http://megapro.kubsu.ru/MegaPro/Web
2. Электронная библиотека трудов ученых КубГУ http://megapro.kubsu.ru/MegaPro/UserEntry?Action=ToDb&idb=6
3. Среда модульного динамического обучения http://moodle.kubsu.ru
4. База учебных планов, учебно-методических комплексов, публикаций и конференций http://infoneeds.kubsu.ru/
5. Библиотека информационных ресурсов кафедры информационных образовательных технологий http://mschool.kubsu.ru;
6. Электронный архив документов КубГУ http://docspace.kubsu.ru/
7. Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ" http://icdau.kubsu.ru/

5.4 Публикации конференций А*

1. Farzana Ahamed Bhuiyan and Akond Rahman. 2021. Characterizing co-located insecure coding patterns in infrastructure as code scripts. In Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE '20). Association for Computing Machinery, New York, NY, USA, 27–32. **https://doi.org/10.1145/3417113.3422154**
2. Michael Hilton, Timothy Tunnell, Kai Huang, Darko Marinov, and Danny Dig. 2016. Usage, costs, and benefits of continuous integration in open-source projects. In Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering (ASE '16). Association for Computing Machinery, New York, NY, USA, 426–437. https://doi.org/10.1145/2970276.2970358
3. Big Data Research (Elsevier) – публикации по анализу, управлению и визуализации данных.
4. Data Science Journal (CODATA) – междисциплинарные исследования данных.
5. ACM Transactions on Knowledge Discovery from Data (TKDD) – методы извлечения знаний из больших данных.
6. https://openreview.net/forum?id=FMMF1a9ifL
7. https://openreview.net/forum?id=ElUrNM9U8c#discussion
8. https://openreview.net/forum?id=JoO6mtCLHD
9. https://aclanthology.org/2024.findings-emnlp.760/
10. https://aclanthology.org/2020.coling-main.588/
11. https://link.springer.com/chapter/10.1007/978-3-030-72113-8_30
12. https://link.springer.com/chapter/10.1007/978-3-031-42448-9_10
13. https://aclanthology.org/2024.findings-naacl.288/

6. Методические указания для обучающихся по освоению дисциплины

6.1 Рекомендации по организации обучения

Освоение дисциплины «Подготовка данных машинного обучения» требует системного подхода и активной самостоятельной работы. По курсу предусмотрено проведение лекционных занятий, на которых дается систематизированный материал по дисциплине. В ходе лекций рассматриваются ключевые концепции. После каждой лекции рекомендуется выполнение практических заданий для закрепления ключевых понятий и методов и самостоятельная работа с дополнительным материалом и литературой.

Лабораторные занятия курса посвящены практическому освоению методов дисциплины «Подготовка данных машинного обучения». На занятиях студенты реализуют задачи сбора и предварительный подготовки данных, finetuning и балансировку данных, в том числе в облачных средах, предоставленных партнерами.

При самостоятельной работе студентам необходимо изучать рекомендованную литературу в виде официальной документации к используемым открытым программным продуктам, облачным платформам.

6.2 Стратегии выполнения лабораторных работ

После изучения базовых концепций рекомендуется выполнение лабораторных работ по схеме:

1. Подготовка данных (нормализация, кодирование, обработка пропусков и выбросов).
2. Feature Engineering (создание новых признаков, отбор признаков).
3. Балансировка данных и аугментация (для изображений и текстов).
4. Построение пайплайнов предобработки.
5. Эксперименты с моделями на подготовленных данных.
6. Анализ результатов и интерпретация моделей.

Пайплайн на python:

```
def run_complete_analysis(dataset, target_column, problem_type='classification'):
    """Полный анализ от данных до интерпретации"""

    # 1. Подготовка данных
    df = load_dataset(dataset)
    X_train, X_test, y_train, y_test = prepare_data(df, target_column)

    # 2. Эксперименты с моделями
    results = run_model_experiments(X_train, X_test, y_train, y_test, problem_type)

    # 3. Анализ результатов
    visualize_results(results, y_test)

    # 4. Интерпретация лучшей модели
    best_model_name = max(results, key=lambda x: results[x]['f1_score'])
    best_model = results[best_model_name]['model']

    interpret_model(best_model, X_train, X_test)

    return results
```

6.3. Рекомендации для студентов с ОВЗ

– Материалы предоставляются в адаптированных форматах: аудиоформат, электронные документы с увеличенным шрифтом.
– Консультации проводятся индивидуально (включая онлайн-формат).
– Лабораторные работы могут быть скорректированы (упрощенные датасеты, расширенные сроки сдачи).

Подход, определяющий установление соответствия кейсов ИП и УГТ (5-7), позволяет четко соотносить этапы развития технологии с вовлеченностью партнера и снижать риски при переходе от лабораторных испытаний к промышленному внедрению.

Кейсы ПАО «Сбербанк»

1. Генеративный ИИ для автоматического составления инвестиционных обзоров

Описание:
Аналитики Сбера ежедневно составляют десятки аналитических и инвестиционных обзоров по рынкам, компаниям, макроэкономике. Задача — исследовать применение LLM для генерации кратких сводок и аналитических отчетов на основе входных данных: биржевые котировки, макроэкономические показатели, рыночные события.

Цель:
Разработать инструмент, способный по структурированным данным и краткому описанию формировать инвестиционный обзор в деловом стиле.

Ожидаемый результат:
Модель, генерирующая аналитические тексты длиной 500–1000 слов с разделами «обзор событий», «рекомендации», «прогнозы», оформленные в формате банка.

2. НЛП-анализ жалоб клиентов в свободной форме

Описание:
В рамках клиентского сервиса Сбербанк обрабатывает обращения из чатов, мобильного приложения и жалобной формы. Требуется построить модель семантического анализа, выделяющую суть обращения, определяющую тональность и потенциальную серьёзность инцидента.

Цель:
Автоматизировать классификацию обращений для ускорения маршрутизации и выявления повторяющихся болевых точек в продуктах и процессах.

Ожидаемый результат:
Прототип модели, автоматически выделяющей темы жалоб (например, «ошибка в приложении», «двойное списание»), их эмоциональную окраску и критичность.

3. Генерация сценариев фишинговых писем для обучения сотрудников

Описание:
Банк проводит киберучения, включая рассылку тестовых фишинговых писем сотрудникам для повышения их устойчивости к социальным атакам. Проект предполагает использование генеративной модели для создания реалистичных фишинговых писем различных типов (поддельные счета, HR-запросы, ИТ-поддержка).

Цель:
Создать генератор, способный на основе заданных параметров (тема, стиль, уровень угрозы) создавать тексты фишинга для тренировок.

Ожидаемый результат:
Набор разнообразных примеров фишинга и оценка их эффективности по реакции сотрудников, а также классификация моделей угроз.

4. Мультимодальный ассистент для банковских отделений

Описание:
Физические отделения Сбербанка внедряют интерактивных консультантов. Предполагается создание мультимодального ИИ-ассистента, который воспринимает речь и визуально ориентируется в пространстве (распознаёт клиента, документы, банкоматы), а также отвечает голосом.

Цель:
Разработать базовый прототип, имитирующий функциональность помощника: ответы на типовые запросы, визуальные подсказки, навигация по отделению.

Ожидаемый результат:
Интерактивная модель, объединяющая голосовой ввод, зрительное восприятие (например, QR-код паспорта), текстовый вывод и жестовую реакцию.

5. Объяснимость и контроль генеративных моделей в банковском ИИ

Описание:
Банк активно использует LLM и NLP-сервисы (в чат-ботах, генерации шаблонов ответов, автоответах на e-mail), однако встает вопрос: как объяснять и контролировать поведение таких моделей, особенно в юридически значимых коммуникациях?

Цель:
Исследовать подходы к трассировке решений LLM (например, через логирование reasoning chain, пост-фильтрацию ответов, встроенные правила).

Ожидаемый результат:
Концепция системы explainability + compliance-модуля, обеспечивающего соответствие генерации стандартам банка и регулятора.

6. Генерация пользовательских сценариев работы в мобильном приложении

Описание:
Банк хочет использовать генеративный ИИ для быстрой симуляции пользовательских сценариев — например, как клиент оформляет вклад, переводит средства, получает уведомление о риске мошенничества.

Цель:
Разработать генератор пошаговых сценариев пользовательского поведения с вариативностью (молодой клиент, пенсионер, ИП).

Ожидаемый результат:
Набор автоматически сгенерированных UX-сценариев, оформленных в виде сценариев для QA или UX-исследований, с логикой действий и типичными ошибками пользователя.

7. Генерация synthetic data для банковских моделей

Описание:
Модели в Сбере требуют большого объёма транзакционных и клиентских данных, которые нельзя использовать напрямую из-за требований ЦБ и ФЗ-152. Задача — разработать метод генерации синтетических банковских данных, максимально близких к реальным по распределениям и поведению.

Цель:
Создать безопасный pipeline генерации данных (например, транзакций, профилей клиентов, шаблонов расходов) для обучения моделей.

Ожидаемый результат:
Синтетический датасет и отчет о метриках приближённости к реальному (TSNE, K-L divergence и др.), с оценкой пригодности для обучения скоринговых или антифрод-моделей.

8. Модель анализа инвестиционной привлекательности малого бизнеса

Описание:
Банк активно развивает кредитование и инвестиционные инструменты для малого и среднего предпринимательства (МСП). Требуется создать модель, которая на основе открытых и банковских данных (выручка, расходы, тип деятельности, отзывы, онлайн-активность) оценивает инвестиционную привлекательность МСП.

Цель:
Разработать систему рейтинговой оценки компаний малого бизнеса с возможностью визуализации факторов и динамики показателей.

Ожидаемый результат:
Модель, присваивающая компании инвестиционный рейтинг (например, A–E), объясняющая ключевые параметры и дающая рекомендации для инвестора.

9. Индивидуальная оценка кредитоспособности клиента на основе поведенческих данных

Описание:
Современный кредитный скоринг выходит за рамки финансовых данных. Необходимо исследовать, как поведенческие и цифровые следы (частота входа в мобильный банк, способы оплаты, география, время отклика) влияют на персональную оценку риска.

Цель:
Разработать ML-модель, оценивающую вероятность дефолта по нестандартным поведенческим признакам (возможно — с explainable AI).

Ожидаемый результат:
Прототип скоринговой модели, которая, помимо стандартных данных, учитывает цифровой профиль клиента и объясняет решения (SHAP, LIME и др.).

10. Предиктивная аналитика возврата инвестиций по инфраструктурным проектам

Описание:
В ряде случаев Сбербанк выступает участником/инвестором в региональных инфраструктурных проектах (жилые массивы, дороги, технопарки). Задача — оценить прогнозируемую эффективность вложений с учётом демографии, миграции, экономической активности.

Цель:
Разработать модель, прогнозирующую ROI на горизонте 3–5 лет, используя внешние источники данных: Росстат, ЕГРЮЛ, кадастр, соцмедиа.

Ожидаемый результат:
Аналитическая модель с возможностью геовизуализации и сценарного анализа (рост/спад, госпрограммы, смена трафика и т.п.).

11. Анализ поведения пользователей в экосистеме цифрового рубля

Описание:
Сбербанк участвует в пилотных проектах по внедрению цифрового рубля. Интерес представляет исследование пользовательских паттернов: как изменяются модели потребления, скорости операций, уровень доверия, сравнение с классическим безналом.

Цель:
Построить модель анализа поведения клиентов, участвующих в транзакциях с цифровым рублем: частота, средний чек, контексты.

Ожидаемый результат:
Отчёт и ML-модель, классифицирующая типы пользователей и выявляющая ключевые различия в предпочтениях и барьерах цифровой валюты.

12. Сравнение text2video / text2img моделей

Описание:
Сбербанк заинтересован в сравнении text2video / text2img моделей (открытые модели,

особенно китайские). Задача требует применения облачных ресурсов партнера для машинного обучения. От студентов требуется навык запуска открытых моделей, планирования, структурирования и логирования экспериментов, совместной работы. Задача может быть распараллелена для сравнения множества моделей независимо в группе студентов.

Цель:
Провести сравнение работы актуальных открытых моделей text2video / text2img.

Ожидаемый результат:
Таблица с результатами экспериментов модель / репозиторий / функционал / требования / оценка производительности / X примеров генераций (было/стало), human_eval по принципу арены (какая лучше)

Кейсы от «АВАЛАБ»

1. LLM и RAG для BI-системы Fastboard

Описание:
Для разрабатываемой компанией BI-системы Fastboard требуется разработать интерфейс на естественном языке для построения отчетов на больших массивах данных в ClickHouse. С помощью LLM необходимо классифицировать запросы пользователей на естественном языке и извлекать фактические параметры для дальнейшего вызова веб-сервиса отчетов.

Цель:
Разработать промпты для классификации и обработки запросов пользователей LLM и преобразования их к вызовам типовых отчетов с фактическими параметрами, извлекаемыми из запроса.

Ожидаемый результат:
Инструмент на основе LLM, позволяющий запрашивать данные о продажах.

2. Анализ обращений клиентов и CRM-переписки

Описание:
В службе клиентского сервиса застройщика ежедневно обрабатываются десятки обращений (e-mail, звонки, мессенджеры). Требуется реализовать систему семантического анализа и классификации NLU: выявлять суть обращений, уровень удовлетворенности, отслеживать повторяющиеся запросы.

Цель:
Автоматизировать первичный разбор и маршрутизацию запросов по тематике (сдача объекта, отделка, документы, жалоба и т.д.).

Ожидаемый результат:
Прототип, который выделяет суть обращений и формирует дашборд по текущим «болям» клиентов.

3. Генеративный ИИ для создания проектной документации по ТЗ

Описание:
В рамках проектирования объектов девелоперской компании архитекторы и инженеры тратят значительное время на подготовку текстовой проектной документации (обоснование

решений, пояснительные записки, описания инженерных систем). Задача — исследовать возможность использования LLM для генерации черновиков проектной документации на основе исходных данных: этажность, материалы, климат, назначение, нормы.

Цель:
Разработать прототип текстового генератора, который помогает специалистам быстрее формировать документацию в соответствии с шаблонами и нормативами.

Ожидаемый результат:
Инструмент на основе LLM, создающий логически стройный и нормативно грамотный текст, поддающийся быстрой редактуре инженером.

4. Мультимодальный агент для анализа строительных площадок

Описание:
ООО «АВА ЛАБ» разрабатывает систему для мониторинга строительных объектов. Требуется создать прототип мультимодального ИИ-агента, способного анализировать изображения со стройплощадки (видео/фото), а также принимать голосовые и текстовые запросы (например, «проверь монтаж перекрытия на 5 этаже»).

Цель:
Объединить возможности компьютерного зрения (распознавание стадии строительства, техники, нарушений) и НЛП (понимание запросов, отчётов).

Ожидаемый результат:
Интерактивный агент, который на запрос специалиста может показать нужный участок, прокомментировать прогресс, зафиксировать нарушения.

4. Генерация рекламного контента для жилых комплексов

Описание:
«АВА ГРУПП» регулярно запускает маркетинговые кампании для жилых комплексов. Необходимо исследовать использование диффузионных моделей для генерации изображений (визуализации интерьеров, окрестностей, видов из окон) и LLM — для описаний квартир, преимуществ района, инфраструктуры.

Цель:
Создать инструменты для быстрой генерации продающих материалов без привлечения дизайнеров и копирайтеров на первых этапах.

Ожидаемый результат:
Набор сгенерированных карточек объектов с текстом, изображением и логикой «живого» рекламного сообщения.

6. Генерация документации и шаблонов договоров

Описание:
Юридический департамент регулярно работает с договорами долевого участия, актами приёма-передачи и другими документами. Использование LLM может значительно сократить время на подготовку черновиков — достаточно ввести параметры сделки.

Цель:
Создать систему, которая генерирует адаптированные тексты документов по вводным данным (тип объекта, этаж, площадь, ФИО, сроки и пр.).

Ожидаемый результат:
Генератор документов в формате Word или PDF с автоматической подстановкой параметров и соблюдением юридического стиля.

7. Модель прогнозирования сроков сдачи объектов на основе текстовых и визуальных данных

Описание:
Девелоперская компания ведёт аналитический архив по срокам строительства. С помощью мультимодальных моделей (текстовые отчёты + фото стройки) можно прогнозировать вероятность отклонения от графика сдачи.

Цель:
Разработать модель, которая по текущему статусу объекта (фото, отчёт СМР) оценивает риски задержек.

Ожидаемый результат:
Прототип, который показывает вероятность отклонений и даёт текстовые пояснения (основанные на распознанных признаках — «не завершены фасадные работы», «монтаж инженерии не начат»).

8. Обратная генерация — ИИ-помощник для покупателей квартир

Описание:
Будущие покупатели часто задают типовые вопросы о квартирах, планировках, ипотеке, акциях, сроках. Вместо call-центра предлагается реализовать LLM-бота, который обрабатывает текстовые и голосовые запросы, показывает планировки, ссылается на PDF-документы и может «объяснять» информацию простым языком.

Цель:
Упростить коммуникацию с клиентами на этапе выбора квартиры и повысить качество первичного контакта.

Ожидаемый результат:
Демо-бот, способный отвечать на вопросы о жилом комплексе, ориентируясь в его характеристиках и маркетинговых документах.

КЕЙСЫ ДЛЯ ООО «СвязьРесурс-Кубань»

Описание:

Компания ООО "СвязьРесурс-Кубань" оказывает услуги связи. Работа с клиентами автоматизирована на базе CRM Битрикс 24. Для компании актуальны вопросы разработки первоначальных версий документов с помощью LLM и в перспективе автоматизации генерации большого количества документов по шаблонам с помощью LLM и RAG системы с интеграцией с Битрикс 24. Задачи включают в себя:

1. Разработка библиотеки промптов для генерации регламентов описания бизнес-процессов Битрикс 24.
2. Разработка библиотеки промптов для генерации техзаданий на основе параметров оказания услуг.
3. Разработка библиотеки промптов для генерации коммерческих предложений на основе параметров оказания услуг.
4. Разработка библиотеки промптов для генерации скриптов работы технической поддержки.
5. Разработка библиотеки промптов для генерации скриптов работы отдела продаж.

6. Апробация и сравнение различных языковых моделей для решения задач.

Цель:

Автоматизировать работу сотрудников по составлению типовых документов.

Ожидаемый результат:

Библиотека промптов и рекомендации по использованию LLM для решения поставленных задач.

7. Материально-техническое обеспечение по дисциплине (модулю)

1. Облачные платформы и сервисы

cloud.ru,YandexCloud,AWS/GCP/Azure– облачные вычисления

2. Системы управления версиями и коллаборации

Git/GitHub/GitLab – контроль версий кода и совместная разработка

2. Свободное ПО (Open Source)

GitLab, GIT, MLFlow, Docker, Kubernetes, Terraform.

Виртуальные машины, кластер Managed Kubernetes и ресурсы GPU в облаке предоставляется индустриальным партнером ПАО «Сбербанк»:

№	Продукт	Параметры продукта	Кол-во	Кол-во конфигураций	Ед. изм.
1	Виртуальная машина	Виртуальная машина 10% vCPU 2 vCPU 4 RAM	1	60	Шт
		ОС Ubuntu 22.04	1		Шт
		Системный диск SSD	1		Шт
			10		Гб
		Аренда публичного IP	1		Шт
2	Виртуальная машина с GPU	Виртуальная машина с GPU NVIDIA® Tesla® V100 2 GPU 8 vCPU 128 ГБ RAM	1	1	Шт
		ОС Ubuntu_24.04	1		Шт
		Системный диск SSD	1		Шт
			2000		Гб
		Диск SSD	1		Шт
			4096		Гб
		Диск SSD	1		Шт
			4096		Гб
		Аренда публичного IP	1		Шт
3	K8S	Master node 8 vCPU 16 RAM	1	1	Шт
		Worker node 10% доля 4 vCPU 32 RAM	5		Шт
		Worker node SSD-NVME	64		Гб
		Аренда публичного IP	1		Шт
4	ML Inference Instance Type GPU	Время работы в месяц	40	1	Ч

		Инстанс 8 x NVIDIA® H100 NVLink PCIe 160 vCPU 1520 GB RAM	1		Шт
		Количество запросов к ML-моделям	1		Млн. Шт
		Кэш ML-моделей	160		Гб
5	LLM	Токены GigaChat 2 Max	50		Млн. Шт
		Токены Embeddings	400		Млн. Шт

Дополнительные облачные ресурсы предоставляются технологическим партнером Yandex Cloud.

№	Вид работ	Наименование учебной аудитории, ее оснащенность оборудованием и техническими средствами обучения
1	Лекционные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения
2	Лабораторные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, проектором, программным обеспечением
3	Практические занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения
4	Групповые (индивидуальные) консультации	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
5	Текущий контроль, промежуточная аттестация	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
6	Самостоятельная работа	Кабинет для самостоятельной работы, оснащенный компьютерной техникой с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в электронную информационно-образовательную среду университета.

Примечание: Конкретизация аудиторий и их оснащение определяется ОПОП.