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 (МОДУЛЯ)

Б1. В.ДВ.01.02 Data-Centric Machine Learning

Направление подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Профиль Современные методы машинного обучения и компьютерного зрения

Форма обучения очная

Квалификация бакалавр

Краснодар 2025

Рабочая программа дисциплины Data-Centric Machine Learning составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии

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 кафедры Вычислительных технологий протокол № 1 « 26 » _августа_ 2025 г.

И.о. заведующего кафедрой (разработчика) Приходько Т.А _____

подпись

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 «Data-Centric Machine Learning » 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 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 «Data-Centric Machine Learning» 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 «Машинное обучение», «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 анализа распределений данных, поиска аномалий и выбросов, статистические тесты для оценки качества данных (нормальность, стационарность), принципы анализа мультимодальных и неструктурированных данных.

Умеет: выбирать методы EDA в зависимости от типа данных и задачи ML, интерпретировать результаты анализа для формулирования гипотез о качестве данных, адаптировать стандартные методы анализа под специфику доменной области.

Владеет библиотеками: pandas-profiling, sweetviz, dataprep, умеет строить интерактивные дашборды для анализа данных. Автоматизирует пайплайны предварительного анализа.

D-1.2 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ.

Знает: Методы формирования и проверки статистических гипотез, принципы A/B тестирования для оценки качества данных, метрики качества данных для ML (class imbalance, label noise, feature quality).

Умеет: Формулировать гипотезы о проблемах в данных на основе EDA, планировать эксперименты по оценке влияния качества данных на модель, валидировать гипотезы о смещениях в данных (bias detection).

Владеет: scipy.stats, statsmodels для статистического тестирования, проводит power analysis для экспериментов с данными, автоматизирует проверку гипотез о качестве данных.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных

Знает: алгоритмы линейного и нелинейного снижения размерности (PCA, t-SNE, UMAP), методы отбора признаков на основе важности и корреляций, метрики оценки качества снижения размерности

Умеет: Выбирать метод снижения размерности в зависимости от задачи и типа данных, интерпретировать результаты визуализации для выявления кластеров и аномалий, оценивать информативность признаков после преобразования.

Владеет: sklearn.decomposition, umap-learn, seaborn, умеет создавать интерактивные визуализации с plotly, оптимизирует параметры снижения размерности для интерпретируемости.

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 sklearn.feature_selection, rfpimp, SHAP, умеет проводить рекурсивное исключение признаков, строит матрицы корреляций и анализирует мультиколлинеарность.

BD-2 Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных

BD-2.1 Определяет требования к наборам и качеству данных для решения задач машинного обучения.

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 Работает с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность.

Знает: методы интеграции данных из разнородных источников, принципы валидации и верификации данных, техники обработки пропущенных значений и аномалий.

Умеет: Проектировать ETL/ELT процессы для сбора данных, разрабатывать правила валидации данных (data contracts), выявлять и устранять проблемы согласованности данных.
Владеет: pandas, pyarrow, dask для обработки больших данных,умеет работать с API, базами данных, облачными хранилищами, автоматизирует проверки качества данных в пайплайнах.

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

LLM-2.1 **Понимает принципы fine-tune**

Знает: Архитектуры и принципы работы современных LLM, методы адаптации моделей: full fine-tuning, LoRA, QLoRA, P-tuning, особенности подготовки данных для дообучения генеративных моделей.

Умеет: Выбирать стратегию дообучения в зависимости от задачи и ресурсов, оценивать риски катастрофического забывания и способы его смягчения, планировать эксперименты по оценке эффективности дообучения.
Понимает технические ограничения разных методов fine-tuning, умеет оценивать вычислительные требования для адаптации моделей, анализирует trade-offs между качеством и стоимостью дообучения.

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

Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки. Форматы данных для разных методов дообучения (SFT, DPO, RLHF), принципы построения диалоговых систем и инструктивных данных, методы аугментации и синтеза данных для NLP, проектировать схемы разметки для задач дообучения LLM.
Умеет: формировать сбалансированные и разнообразные обучающие выборки
Оценивать качество и репрезентативность созданных датасетов. Умеет создавать синтетические данные с помощью LLM
Владеет инструментами для разметки текстовых данных (Label Studio, etc.), **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: дескриптивную статистику и методы анализа распределений, статистические тесты для сравнения групп и выявления различий, методы анализа временных рядов и последовательностей.
Умеет: выбирать статистические методы в зависимости от типа данных и гипотез, интерпретировать результаты статистического анализа для принятия решений. формулировать выводы на основе статистических доказательств.
Владеет: scipy.stats, statsmodels, pingouin, умеет проводить анализ мощности (power analysis), визуализирует результаты статистического анализа.

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):		**34,2**	**34,2**
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 Data-Centric Machine Learning. Процесс разметки данных (Data Labeling)	8	2		2	4
2.	Стратегии разметки данных. Согласованность разметки и разрешение конфликтов	8	2		2	4
3.	Поиск и очистка от шума в данных (Data Cleaning).	8	2		2	4
4.	Обогащение данных: Аугментация и работа с несбалансированными данными (Class Imbalance)	8	2		2	4
5.	Глубинный анализ ошибок и срезов данных (Error & Slice Analysis)	8	2		2	4
6.	Синтетические данные и управление данными	9	2		2	5
7.	Data-Centric подходы для NLP и CV	8,8	2		2	4,8
8.	Инфраструктура, мониторинг и этика	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 Data-Centric Machine Learning. Процесс разметки данных (Data Labeling)	Почему данные важнее моделей? Противопоставление Model-Centric vs. Data-Centric. **Ключевые темы:** Жизненный цикл данных в ML, инструкции для разметчиков (Labeling Guidelines), управление качеством разметки, метрики согласованности (Inter-Annotator Agreement).	ЛР
2.	Стратегии разметки данных. Согласованность разметки и разрешение конфликтов	Как размечать меньше, но лучше. **Ключевые темы:** Стратегии Active Learning (Uncertainty Sampling). Weak Supervision: создание тренировочных данных с помощью эвристик и правил (Snorkel).	ЛР
3.	Поиск и очистка от шума в данных (Data Cleaning).	Диагностика и лечение "больных" данных. Ключевые темы: Типы шума (в features и labels). Методы обнаружения выбросов и проблемных	ЛР

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
		меток. Практическое использование библиотеки cleanlab.	
4.	Обогащение данных: Аугментация и работа с несбалансированными данными (Class Imbalance)	Как улучшить данные, не собирая новые. Ключевые темы: Методы аугментации для изображений (albumentations) и текста (nlpaug). Техники борьбы с несбалансированностью классов (SMOTE, взвешивание классов).	ЛР
5.	Глубинный анализ ошибок и срезов данных (Error & Slice Analysis)	Понимание того, где и почему модель ошибается. Ключевые темы: Построение таксономии ошибок. Срезовый анализ (Slice-Based Analysis) для выявления слабых мест модели в конкретных подгруппах данных.	ЛР
6.	Синтетические данные и управление данными	Создание данных и управление их жизненным циклом. Ключевые темы: Когда и зачем использовать синтетические данные (GANs, Diffusion). Версионирование данных и моделей с DVC. Понятие о дрейфе данных (Data Drift)	ЛР
7.	Data-Centric подходы для NLP и CV	Особенности работы с текстом и изображениями. Ключевые темы: Специфика разметки NER и детекции объектов. Активное обучение и аугментация для разных модальностей. Инструменты (Label Studio, Doccano, CVAT).	ЛР
8.	Инфраструктура, мониторинг и этика	Фокус: Data-Centric AI в продакшене. Ключевые темы: Платформы для разметки, мониторинг данных (Evidently AI), понятие о смещениях (Data Bias) и методах обеспечения справедливости (Fairness).	ЛР

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 Data-Centric Machine Learning. Процесс разметки данных (Data Labeling)	**Лаб 1: Model-Centric vs. Data-Centric на практике** **Цель:** Прочувствовать разницу. **Задание:** Испортить часть меток в датасете (CIFAR-10/IMDb). Сначала пытаться улучшить модель, меняя гиперпараметры. Затем улучшить качество, исправив метки. Сравнить результаты.	Опрос по теоретическому материалу. Отчет по лабораторной работе.

2.	Стратегии разметки данных. Согласованность разметки и разрешение конфликтов	**Лаб 2: Активное обучение для классификации текста** **Цель:** Научиться эффективно выбирать данные для разметки. **Задание:** Реализовать цикл Active Learning (стратегия Uncertainty Sampling) для текстового классификатора. Визуализировать, как растет точность с количеством запрошенных меток.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
3.	Поиск и очистка от шума в данных (Data Cleaning).	**Лаб 3: Очистка датасета с помощью cleanlab** **Цель:** Найти и исправить ошибки в разметке. **Задание:** Намеренно зашумлить метки в датасете. Обучить модель, использовать cleanlab для поиска потенциально неверных меток. Проверить их, исправить и переобучить модель. Зафиксировать прирост качества.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
4.	Обогащение данных: Аугментация и работа с несбалансированными данными (Class Imbalance)	**Лаб 4: Аугментация и борьба с дисбалансом** **Цель:** Увеличить разнообразие данных и решить проблему дисбаланса. **Задание:** Для изображений: использовать albumentations для создания пайплайна аугментаций. Для текста: использовать nlpaug для аугментации. Применить SMOTE или взвешивание классов к несбалансированному датасету. Оценить влияние на качество модели.	Контрольная работа №2 Проверка выполнения домашних работ.
5.	Глубинный анализ ошибок и срезов данных (Error & Slice Analysis)	**Лаб 5: Анализ ошибок и срезов данных** **Цель:** Научиться диагностировать слабые места модели. **Задание:** Натренировать модель и провести детальный анализ её ошибок. Построить Confusion Matrix. Определить 2-3 среза данных (например, "короткие тексты", "изображения с темным фоном"), на которых модель работает хуже, и предложить гипотезы, почему.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
6.	Синтетические данные и управление данными	**Лаб 6: Data-Centric пайплайн для NLP (NER)** **Цель:** Применить ключевые техники к текстовым данным. **Задание:** Настроить проект в Label Studio/Doccano для разметки NER. Провести разметку, применить аугментацию текста, обучить модель (например, spaCy) и провести анализ ошибок.	Опрос по теоретическому материалу. Отчет по лабораторной работе.

7.	Data-Centric подходы для NLP и CV	**Лаб 7: Синтетические данные и Weak Supervision** **Цель:** Освоить методы генерации и "шумной" разметки данных. **Задание:** . **Синтетика:** Сгенерировать синтетические изображения с помощью предобученной GAN и добавить их в тренировочный набор. **Weak Supervision:** Использовать Snorkel для создания Labeling Functions и генерации тренировочных данных для текстовой классификации без ручной разметки.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
8.	Инфраструктура, мониторинг и этика. Итоговый проект.	**Лаб 8: Финальный проект: Сквозной Data-Centric пайплайн** **Цель:** Интегрировать все изученные методы. **Задание:** Выбрать датасет. Реализовать пайплайн, включающий: 1. Начальную разметку/активное обучение. 2. Очистку данных (cleanlab). 3. Аугментацию. Анализ ошибок и целенаправленное улучшение слабых срезов. Представить отчет с метриками до и после каждого шага.	Опрос по теоретическому материалу. Отчет по лабораторной работе.

2.3.4 Примерная тематика курсовых работ (проектов)

Курсовая работа не предусмотрена. В качестве курсового проекта студенты защищают датацентрические методы обработки данных для 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	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.

	занятиям	
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:

1. Технология коммуникативного обучения – направлена на формирование коммуникативной компетентности студентов, которая является базовой, необходимой для адаптации к современным условиям межкультурной коммуникации.
2. 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.

3. Технология модульного обучения – предусматривает деление содержания дисциплины на достаточно автономные разделы (модули), интегрированные в общий курс.
4. 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:
5. Технология использования компьютерных программ – позволяет эффективно дополнить процесс обучения языку на всех уровнях.
6. Интернет-технологии – предоставляют широкие возможности для поиска информации, разработки научных проектов, ведения научных исследований.
7. Технология индивидуализации обучения – помогает реализовывать личностно-ориентированный подход, учитывая индивидуальные особенности и потребности учащихся.
8. Проектная технология – ориентирована на моделирование социального взаимодействия учащихся с целью решения задачи, которая определяется в рамках профессиональной подготовки, выделяя ту или иную предметную область.
9. Технология обучения в сотрудничестве – реализует идею взаимного обучения, осуществляя как индивидуальную, так и коллективную ответственность за решение учебных задач.
10. Игровая технология – позволяет развивать навыки рассмотрения ряда возможных способов решения проблем, активизируя мышление студентов и раскрывая личностный потенциал каждого учащегося.
11. Технология развития критического мышления – способствует формированию разносторонней личности, способной критически относиться к информации, умению отбирать информацию для решения поставленной задачи.
12. 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:

13. 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;
14. проектная технология - индивидуальная или коллективная деятельность по отбору, распределению и систематизации материала по определенной теме, в результате которой составляется проект;
15. анализ конкретных ситуаций - анализ реальных проблемных ситуаций, имевших место в соответствующей области профессиональной деятельности, и поиск вариантов лучших решений;
16. 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 «Data-Centric Machine Learning ».

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 «Data-Centric Machine Learning ».

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 Data-Centric Machine Learning. Процесс разметки данных (Data Labeling)	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №1*	*Вопросы к зачету*
2	Стратегии разметки данных. Согласованность разметки и разрешение конфликтов	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №2*	*Вопросы к зачету*
3	Поиск и очистка от шума в данных (Data Cleaning).	BD-1, BD-2, ML-2	*Лабораторная работа №3*	*Вопросы к зачету*
4	Обогащение данных: Аугментация и работа с несбалансированными данными (Class Imbalance)	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №4*	*Вопросы к зачету*
5	Глубинный анализ ошибок и срезов данных (Error & Slice Analysis)	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №5*	*Вопросы к зачету*
6	Синтетические данные и управление данными	BD-1, BD-2, LLM-2, MF-4, ML-2	*Лабораторная работа №6*	*Вопросы к зачету*
7	Data-Centric подходы для NLP и CV	LLM-2, MF-4, ML-2	*Лабораторная работа №7*	*Вопросы к зачету*
8	Инфраструктура, мониторинг и этика. Итоговый проект.	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 анализа распределений данных, поиска аномалий и выбросов, статистические тесты для оценки качества данных (нормальность, стационарность), принципы анализа мультимодальных и неструктурированных данных.

Умеет: выбирать методы EDA в зависимости от типа данных и задачи ML, интерпретировать результаты анализа для формулирования гипотез о качестве данных, адаптировать стандартные методы анализа под специфику доменной области.

Владеет библиотеками: pandas-profiling, sweetviz, dataprep, умеет строить интерактивные дашборды для анализа данных. Автоматизирует пайплайны предварительного анализа.

D-1.2 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ.

Знает: Методы формирования и проверки статистических гипотез, принципы A/B тестирования для оценки качества данных, метрики качества данных для ML (class imbalance, label noise, feature quality).

Умеет: Формулировать гипотезы о проблемах в данных на основе EDA, планировать эксперименты по оценке влияния качества данных на модель, валидировать гипотезы о смещениях в данных (bias detection).

Владеет: scipy.stats, statsmodels для статистического тестирования, проводит power analysis для экспериментов с данными, автоматизирует проверку гипотез о качестве данных.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных

Знает: алгоритмы линейного и нелинейного снижения размерности (PCA, t-SNE, UMAP), методы отбора признаков на основе важности и корреляций, метрики оценки качества снижения размерности

Умеет: Выбирать метод снижения размерности в зависимости от задачи и типа данных, интерпретировать результаты визуализации для выявления кластеров и аномалий, оценивать информативность признаков после преобразования.

Владеет: sklearn.decomposition, umap-learn, seaborn, умеет создавать интерактивные визуализации с plotly, оптимизирует параметры снижения размерности для интерпретируемости.

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 sklearn.feature_selection, rfpimp, SHAP, умеет проводить рекурсивное исключение признаков, строит матрицы корреляций и анализирует мультиколлинеарность.

BD-2 **Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных**

BD-2.1 Определяет требования к наборам и качеству данных для решения задач машинного обучения.

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 Работает с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность.

Знает: методы интеграции данных из разнородных источников, принципы валидации и верификации данных, техники обработки пропущенных значений и аномалий.

Умеет: Проектировать ETL/ELT процессы для сбора данных, разрабатывать правила валидации данных (data contracts), выявлять и устранять проблемы согласованности данных.

Владеет: pandas, pyarrow, dask для обработки больших данных,умеет работать с API, базами данных, облачными хранилищами, автоматизирует проверки качества данных в пайплайнах.

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

LLM-2.1 **Понимает принципы fine-tune**

Знает: Архитектуры и принципы работы современных LLM, методы адаптации моделей: full fine-tuning, LoRA, QLoRA, P-tuning, особенности подготовки данных для дообучения генеративных моделей.

Умеет: Выбирать стратегию дообучения в зависимости от задачи и ресурсов, оценивать риски катастрофического забывания и способы его смягчения, планировать эксперименты по оценке эффективности дообучения.

Понимает технические ограничения разных методов fine-tuning, умеет оценивать вычислительные требования для адаптации моделей, анализирует trade-offs между качеством и стоимостью дообучения.

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

Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки. Форматы данных для разных методов дообучения (SFT, DPO, RLHF), принципы построения диалоговых систем и инструктивных данных, методы аугментации и синтеза данных для NLP, проектировать схемы разметки для задач дообучения LLM.

Умеет: формировать сбалансированные и разнообразные обучающие выборки

Оценивать качество и репрезентативность созданных датасетов. Умеет создавать синтетические данные с помощью LLM

Владеет инструментами для разметки текстовых данных (Label Studio, etc.), **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: дескриптивную статистику и методы анализа распределений, статистические тесты для сравнения групп и выявления различий, методы анализа временных рядов и последовательностей.

Умеет: выбирать статистические методы в зависимости от типа данных и гипотез, интерпретировать результаты статистического анализа для принятия решений. формулировать выводы на основе статистических доказательств.

Владеет: scipy.stats, statsmodels, pingouin, умеет проводить анализ мощности (power analysis), визуализирует результаты статистического анализа.

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: Сравнение Model-Centric vs Data-Centric подходов

Цель: На практике ощутить разницу между подходами
Соответствует: BD-1.1, BD-1.2, ML-2.1, ML-2.3

Датасет: CIFAR-10 (Canadian Institute For Advanced Research) - один из самых популярных бенчмарков для компьютерного зрения.
Основные параметры:

- Размер: 60,000 цветных изображений 32×32 пикселя
- Классы: 10 категорий, по 6,000 изображений каждый
- Разделение: 50,000 тренировочных + 10,000 тестовых
- Цвет: RGB (3 канала)

Задание:

1. Model-Centric фаза:
   - Возьмите датасет CIFAR-10 с 30% зашумленными метками;
   - Обучите 3 разные архитектуры (Simple CNN, ResNet-18, EfficientNet-B0);
   - Для каждой пробуйте разные гиперпараметры (learning rate, оптимизаторы);
   - Зафиксируйте лучший результат.
2. Data-Centric фаза:
   - Используйте ту же Simple CNN
   - Примените cleanlab для поиска проблемных меток
   - Вручную проверьте и исправьте 100 самых "шумных" примеров
   - Добавьте базовую аугментацию данных
2. Сравнение:
   - Постройте таблицу результатов.
   - Сравните затраченное время и улучшение accuracy.

Выходные артефакты:

- Jupyter notebook с полным пайплайном;
- Таблица сравнения метрик;
- Анализ: какие методы дали наибольший прирост.

Лабораторная работа 2: Активное обучение для текстовой классификации

Цель: Освоить стратегии эффективной разметки данных
Соответствует: BD-2.1, BD-2.2, LLM-2.2
Задание:

1. Настройка окружения:

- Установите Label Studio;
- Подготовьте датасет отзывов IMDb (10к немаркированных примеров).

2. Реализация активного обучения:

python

```
from modAL.models import ActiveLearner
from sklearn.ensemble import RandomForestClassifier

# Начальная выборка - 100 случайных примеров
learner = ActiveLearner(
   estimator=RandomForestClassifier(),
   X_training=X_initial, y_training=y_initial )
# Стратегия запроса - uncertainty sampling
for idx in range(10):  # 10 итераций
   query_idx, query_inst = learner.query(X_pool, n_instances=20)
   # Разметить выбранные примеры в Label Studio
   learner.teach(X_pool[query_idx], y_new_labels)
```

3. Сравнение стратегий:

- Uncertainty Sampling (entropy)
- Diversity Sampling
- Random Sampling (baseline)

Выходные артефакты:

- Learning curves для разных стратегий.
- Экспортированный размеченный датасет из Label Studio.
- Анализ экономии на разметке.

Лабораторная работа 3: Глубинный анализ ошибок и срезов данных

Цель: Научиться диагностировать слабые места модели
Соответствует: BD-1.1, BD-1.2, MF-4.3, ML-2.3
Задание:

1. Базовое обучение:

- Обучите ResNet-50 на датасете животных (10 классов)
- Получите baseline accuracy = 85%

2. Анализ ошибок:

python

```
# Построение таксономии ошибок
error_categories = {
   'similar_classes': ['cat', 'lynx', 'tiger'],
   'low_quality': ['blurry', 'occluded', 'dark'],
   'background_confusion': ['animal vs background']
}
```

```
# Анализ срезов
from sliceline import slicefinder
slices = slicefinder.find_slices(model, X_test, y_test)
```

3. Целевое улучшение:

   Для худшего среза ("темные изображения кошек"):

- Добавьте аугментацию (brightness adjustment)
- Соберите дополнительные примеры
- Переобучите модель

Выходные артефакты:

- Confusion Matrix с аннотациями.
- Dashboard с визуализацией срезов.
- Отчет по улучшению проблемных категорий.

Лабораторная работа 4: Data-Centric пайплайн для NLP

Цель: Построить полный пайплайн для текстовых данных
Соответствует: BD-1.4, LLM-2.1, LLM-2.2, ML-2.2
Задание:

1. Разметка NER:

- Настройте Doccano для разметки именованных сущностей
- Разметьте 200 примеров из доменной области (медицина/юриспруденция)

2. Weak Supervision:

```
python
from snorkel.labeling import labeling_function

@labeling_function()
def medical_terms(x):
   med_terms = ['patient', 'diagnosis', 'treatment', 'symptoms']
   return MEDICAL if any(term in x.text.lower() for term in med_terms) else ABSTAIN

# Применение правил
from snorkel.labeling import PandasLFApplier
lfs = [medical_terms, drug_mentions, ...]
applier = PandasLFApplier(lfs)
L_train = applier.apply(df_train)
```

3. Дообучение модели:

- Fine-tune BERT на размеченных данных
- Сравните качество с baseline

Выходные артефакты:

- Размеченный датасет в Doccano;
- Модель с улучшенными метриками на целевом домене;
- Анализ эффективности weak supervision.

Лабораторная работа 5: Работа с несбалансированными данными

Цель: Освоить методы борьбы с дисбалансом классов
Соответствует: ML-2.3, BD-1.2, MF-4.1

Задание:

1. Анализ дисбаланса:

- Возьмите датасет мошеннических транзакций (Fraud Detection)
- Проанализируйте распределение классов (99.5% vs 0.5%)

2. Применение методов:

python

```
# SMOTE
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)

# Взвешивание классов
from sklearn.utils.class_weight import compute_class_weight
class_weights = compute_class_weight('balanced', classes=np.unique(y_train), y=y_train)
```

3. Оценка эффективности:

- Сравните метрики: Accuracy, Precision, Recall, F1, PR-AUC
- Проанализируйте бизнес-impact разных подходов

Выходные артефакты:

- Сравнительная таблица метрик
- ROC и Precision-Recall кривые
- Рекомендации по выбору метода для конкретного кейса

Лабораторная работа 6: Поиск и исправление ошибок в данных с cleanlab

Цель: Освоить автоматизированные методы контроля качества данных
Соответствует: BD-1.1, BD-2.2, MF-4.1
Задание:

1. Создание зашумленного датасета:

- Возьмите чистый датасет (MNIST)
- Намеренно испортите 20% меток

2. Обнаружение ошибок:

python

```
import cleanlab
from cleanlab.classification import CleanLearning

# Находим проблемные примеры
label_issues = cleanlab.filter.find_label_issues(
   labels=y_train_noisy,
   pred_probs=pred_probs,  # предсказания модели
   return_indices_ranked_by='self_confidence'
)

# Визуализация самых проблемных
worst_issues = label_issues[:50]
```

3. Исправление и переобучение:

- Вручную проверьте топ-50 проблемных примеров;
- Исправьте метки и переобучите модель;
- Измерьте прирост качества.

Выходные артефакты:

- Визуализация исправленных примеров;
- График улучшения accuracy после исправлений;
- Скрипт для автоматического поиска ошибок.

Лабораторная работа 7: Синтетические данные и аугментация

Цель: Освоить методы генерации и обогащения данных
Соответствует: BD-2.1, LLM-2.2, ML-2.2
Задание:

1. Аугментация изображений:

python
```
import albumentations as A

transform = A.Compose([
   A.Rotate(limit=30, p=0.5),
   A.RandomBrightnessContrast(p=0.2),
   A.Cutout(num_holes=8, max_h_size=8, max_w_size=8, p=0.3),
])

# Применение к датасету
augmented_images = [transform(image=img)['image'] for img in original_images]
```

2. Генерация синтетических данных:

- Используйте предобученную GAN (StyleGAN) для генерации лиц;
- Сгенерируйте 1000 синтетических изображений для миноритарного класса.

3. Синтетические текстовые данные:

- Используйте GPT для генерации текстовых примеров;
- Примените back-translation для аугментации.

Выходные артефакты:

- Увеличенный датасет с синтетическими примерами;
- Сравнение качества моделей на оригинальных и обогащенных данных;
- Анализ diversity синтетических данных.

Лабораторная работа 8: Сквозной Data-Centric пайплайн

Цель: Интегрировать все изученные методы в единый пайплайн
Соответствует: Все компетенции
Задание:
Реализуйте полный пайплайн для реальной задачи (на выбор):

- Классификация медицинских изображений
- Детекция спама в email
- Прогнозирование оттока клиентов

Этапы:

1. Анализ и разметка:

- Active Learning для начальной разметки;
- Weak Supervision для масштабирования.

2. Очистка и обогащение:

- cleanlab для поиска ошибок
- Аугментация и синтетические данные;
- Балансировка классов.

3. Анализ и улучшение:

- Error Analysis и Slice Analysis
- Целевое улучшение проблемных срезов;
- Статистическая валидация улучшений.

Выходные артефакты:

- Полный Jupyter notebook с пайплайном;
- Дашборд с метриками на каждом этапе;
- Финальный отчет с анализом влияния каждого Data-Centric метода на итоговое качество.

Каждая работа включает:

- Четкие критерии оценки;
- Примеры кода для старта;
- Шаблоны для отчетности;
- Ссылки на датасеты и инструменты.

Это позволяет студентам системно осваивать Data-Centric подход на практике.

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.

4.3. Примеры контрольных заданий для промежуточной аттестации и на зачет

Практический блок (на компьютере)
Задание 1: Анализ и очистка данных (BD-1.2, BD-2.2)
python
Дан зашумленный датасет CIFAR-10 с 25% ошибок в метках
Ваши задачи:
1. Проведите EDA: распределение классов, визуализация примеров;
2. Используйте cleanlab для поиска потенциально неверных меток;
3. Исправьте топ-50 самых проблемных примеров;
4. Обучите модель до и после очистки, сравните ac;curacy;
5. Проанализируйте, какие типы ошибок были исправлены.

Критерии оценки:

✓ Качество EDA и визуализации (2 балла)
✓ Корректное использование cleanlab (2 балла)
✓ Сравнительный анализ (1 балл)
Макс: 5 баллов

Задание 1: Активное обучение (BD-2.1, LLM-2.2)

python
Дан немаркированный датасет текстовых отзывов (10,000 примеров)
Изначально размечено только 100 случайных примеров
Ваши задачи:
1. Реализуйте стратегию активного обучения с uncertainty sampling;
2. Проведите 5 итераций, запрашивая по 20 меток за итерацию;
3. Постройте learning curve (качество от объема размеченных данных);
4. Сравните с random sampling baseline;
5. Проанализируйте, какие примеры выбирались для разметки.

Критерии оценки:

✓ Рабочая реализация AL (2 балла)
✓ Корректные эксперименты (2 балла)
✓ Анализ результатов (1 балл)
Макс: 5 баллов

Зачетные задания (итоговые)

Комплексный теоретический экзамен

Билет №1

1. Опишите полный пайплайн Data-Centric подхода для задачи медицинской диагностики по изображениям, начиная от сбора данных до мониторинга в продакшене.
2. Какие этические риски возникают при работе с медицинскими данными и как их mitigate с помощью Data-Centric методов?
3. Предложите стратегию разметки, учитывая необходимость привлечения экспертов-врачей

Билет №2

1. Объясните, как методы Weak Supervision могут ускорить разметку данных для задачи NER в юридических документах. Приведите конкретные примеры LF.
2. Как оценить качество слаборазмеченных данных и когда их использование оправдано?
3. Опишите процесс интеграции Weak Supervision в промышленный ML-пайплайн.

Зачетно-экзаменационные материалы для промежуточной аттестации (зачет)

Теоретические вопросы

1. Основные концепции Data-Centric AI

1. В чем принципиальное отличие Model-Centric и Data-Centric подходов в машинном обучении?
2. Назовите 3 ключевых преимущества Data-Centric подхода и приведите примеры.
3. Что такое "гипотеза неприятного закона" (Unpleasant Law) в контексте качества данных?
4. Опишите жизненный цикл данных в ML-проекте с Data-Centric точки зрения.

2. Разметка и качество данных

5. Какие стратегии активного обучения (Active Learning) вы знаете? В каких случаях применяется каждая?
6. Как оценить согласованность разметчиков? Опишите метрики Cohen's Kappa и Fleiss' Kappa.
7. Что такое Weak Supervision? Приведите примеры использования в реальных задачах.
8. Какие требования к качеству данных необходимо формулировать перед началом ML-проекта?

3. Методы работы с данными

9. Опишите методы обнаружения и очистки данных от шума в признаках и метках.
10. Какие методы аугментации данных наиболее эффективны для изображений и текста?
11. Как бороться с несбалансированностью данных? Сравните методы на уровне данных и алгоритмов.
12. Что такое синтетические данные? Когда их использование оправдано, а когда - нет?

4. Анализ и мониторинг

13. Как проводить глубинный анализ ошибок (Error Analysis)? Что такое таксономия ошибок?
14. Что такое срезовый анализ (Slice Analysis) и для чего он применяется?
15. Опишите методы обнаружения и мониторинга дрейфа данных (Data Drift).
16. Какие статистические методы наиболее полезны для анализа качества данных?

Практические вопросы

5. Инструменты и реализации

17. Как бы вы организовали процесс разметки для задачи NER с использованием Label Studio?
18. Опишите пайплайн использования библиотеки cleanlab для поиска проблемных меток.
19. Как настроить DVC для версионирования данных и моделей в проекте?
20. Какие инструменты вы бы использовали для мониторинга качества данных в продакшене?

6. Решение проблем

21. Ваша модель показывает 95% accuracy, но плохо работает в продакшене. С чего начнете исследование?
22. Как бы вы улучшили качество модели, если нет возможности собирать новые данные?
23. Обнаружено, что 30% меток в тренировочном наборе содержат ошибки. Ваш план действий?
24. Модель хорошо работает на большинстве данных, но полностью проваливается на определенной подгруппе. Что делать?

7. Кейсовые вопросы

25. Опишите, как применить Data-Centric подход для задачи медицинской диагностики по изображениям.
26. Как бы вы построили пайплайн разметки для мультиязычного чат-бота?
27. Предложите стратегию работы с данными для задачи обнаружения мошеннических операций (1% положительных классов).
28. Как обеспечить справедливость (fairness) модели при наличии смещений в данных?

Вопросы на применение знаний

8. Архитектурные и процессные вопросы

29. Как интегрировать Data-Centric практики в существующий ML-пайплайн компании?
30. Какие метрики качества данных следует отслеживать на постоянной основе?
31. Как организовать процесс непрерывного улучшения данных в продакшн-системе?
32. Опишите роли и ответственности в Data-Centric команде.

9. Экономика и менеджмент

33. Как обосновать бизнесу инвестиции в улучшение качества данных?
34. Как оценить ROI от внедрения Data-Centric подходов?
35. Какие риски проекта можно снизить с помощью Data-Centric методов?
36. Как планировать бюджет на разметку и улучшение данных?

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 дата-центричного анализа данных методами 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
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 А*:

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 A*

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 «Data-Centric Machine Learning » 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 «Data-Centric Machine Learning». 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.

Стратегия 1: Поэтапное освоение методов

Принцип: "От простого к сложному"

План выполнения:

1. Подготовительный этап (30 минут)
    - Изучить описание задачи и теоретическую справку
    - Подготовить окружение (установить библиотеки, загрузить датасет)
    - Выполнить базовый EDA
2. Базовая реализация (1 час)
    - Реализовать требуемые методы по шаблону/примеру
    - Получить первые результаты
3. Экспериментальный этап (1 час)
    - Модифицировать параметры методов
    - Сравнить разные подходы
    - Зафиксировать наблюдения
4. Аналитический этап (30 минут)
    - Проанализировать результаты
    - Сформулировать выводы
    - Подготовить отчет

Стратегия 2: Командная работа с ролями

Ролевая модель в команде 3-4 человека:

Data Analyst (Аналитик данных):

- Проводит EDA и анализ качества данных
- Формулирует гипотезы для улучшения
- Визуализирует результаты

Data Engineer (Инженер данных):

- Настраивает пайплайны обработки
- Реализует методы очистки и аугментации
- Обеспечивает воспроизводимость

ML Engineer (ML-инженер):

- Обучает и оценивает модели
- Проводит эксперименты
- Анализирует метрики качества

Пример workflow для команды:
text
```
Неделя 1: Analyst → EDA → Гипотезы
    ↓
Неделя 2: Engineer → Реализация → ML Engineer → Эксперименты
    ↓
Неделя 3: Все вместе → Анализ результатов → Подготовка отчета
```

Стратегия 3: Research-ориентированный подход
Для продвинутых студентов
Принцип: "Не просто сделать, а исследовать"
Методология:

1. Problem Framing (формулировка проблемы)
   - Какую проблему данных решаем?
   - Какие есть гипотезы по улучшению?
2. Experimental Design (дизайн эксперимента)
   - Какие методы сравниваем?
   - Какие метрики используем?
   - Как обеспечиваем достоверность?
3. Iterative Improvement (итеративное улучшение)
   - `Базовый уровень → Метод 1 → Метод 2 → Комбинация`
   - После каждой итерации: анализ, выводы, следующий шаг
4. Generalization Analysis (анализ обобщаемости)
   - На каких данных метод работает хорошо?
   - В каких условиях дает ухудшение?
   - Практические рекомендации

Пример для ЛР 3 (Анализ ошибок):
python
```
# Вместо простого выполнения - исследовательский подход:
research_questions = [
  "Какие типы ошибок преобладают в нашей модели?",
  "Связаны ли ошибки с определенными характеристиками данных?",
  "Какой метод улучшения данных наиболее эффективен для наших ошибок?" ]
```

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.