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.05.01 Анализ данных машинного обучения

Направление подготовки 01.03.02 Прикладная математика и информатика

Профиль Современные методы машинного обучения и компьютерного зрения

Форма обучения очная

Квалификация бакалавр

Краснодар 2025

Рабочая программа дисциплины АНАЛИЗ ДАННЫХ МАШИННОГО ОБУЧЕНИЯ составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 01.03.02 Прикладная математика и информатика

Программу составил(а):

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

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

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

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 «Анализ данных машинного обучения» является формирование у студентов систематизированных знаний, практических умений и навыков применения современных методов искусственного интеллекта, машинного обучения для решения задач анализа данных машинного обучения в различных предметных областях.

Дисциплина направлена на развитие способности выбирать, реализовывать, оценивать и интерпретировать модели классификации.

1.2 Задачи дисциплины

1. Изучение теоретических основ задач машинного обучения;
2. Углубление знаний о современных алгоритмах машинного обучения (логистическая регрессия, SVM, деревья решений, байесовский классификатор, ансамбли, алгоритмы кластерного анализа, алгоритмы снижения размерности);
3. Приобретение практических навыков анализа данных, проектирования, обучения, оценки и оптимизации моделей классификации с использованием современных инструментов (R, Python, scikit-learn, PyTorch/TensorFlow/Keras);
4. Развитие умений анализировать результаты классификации, кластеризации, выбирать метрики качества, интерпретировать работу моделей;
5. Формирование навыков применения методов машинного обучения для решения прикладных задач.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Анализ данных машинного обучения» относится к части, формируемой участниками образовательных отношений Блока 1 "Дисциплины (модули) по выбору" учебного плана (Б1.В.ДВ.05.01).

Дисциплина изучается в 7-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 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
2. Оптимизация производительности и масштабирование моделей
3. Разработка ML-пайплайнов и автоматизация процессов
4. Мониторинг качества моделей в продуктиве
5. Интеграция ML-решений с бизнес-приложениями

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

Задачи:

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

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

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

BD-1	**Способен осуществлять поиск, сбор, очистку и предварительный анализ данных (П)**
BD-1.1	Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи **Знает** методы заполнения пропусков в данных и удаления выбросов в табличных данных (случайные величины) Имеет навыки (**умеет**) очистки зашумленных временных рядов и изображений. Обнаруживает и устраняет выбросы в данных временных рядов. **Владеет** подходами к заполнению пропусков в данных временных рядов и изображений.
BD-1.2	Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ **Знает** основные методы понижения размерности **Умеет** применить основные методы понижения размерности и подбирает оптимальную размерность в зависимости от необходимой доли объяснённой дисперсии. **Владеет** методологией применения существующих библиотек, реализующих методы понижения размерности.
BD-1.3	Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных **Знает** и умеет применить основы методов отбора признаков и выбирает оптимальное подмножество признаков. **Владеет** методологией применения существующих библиотек, реализующих методы отбора признаков.
BD-1.4	**Знает** и умеет применить методы отбора признаков.

Владеет способностью применять методы отбора признаков данных, значимых для исследования.

Умеет отбирать признаки данных, значимые для исследования,

Владеет методами finetuning

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

BD-2.1 Знает, как сформировать требования для набора данных. Владеет умениями по формированию требований к наборам и качеству данных для решения задач машинного обучения

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

BD-2.3 Применяет инструменты и практики непрерывной интеграции данных (DataOps)

Умеет применять инструменты интеграции данных. **Владеет** навыками непрерывной интеграции данных (DataOps)

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

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

Знает: основные подходы к тонкой настройке: полная настройка всех параметров, поэтапная разморозка слоев, методы эффективной тонкой настройки (P-Tuning, LoRA, QLoRA, Adapter). Гиперпараметры, критически важные для fine-tune: learning rate, scheduler, batch size, и их отличия от обучения с нуля.

Умеет: Отличать дообучение от первичного обучения, знает базовые процедуры **fine-tune,** анализировать задачу и выбирать наиболее подходящий метод fine-tune (полная настройка vs. эффективные методы).

Владеет: **Навыком** осознанного выбора стратегии fine-tune под ограничения (вычислительные ресурсы, объем данных, требования к качеству). Применяет fine-tune к предобученным моделям на новых датасетах.

• **Методами** анализа и интерпретации процесса дообучения (использование логов, графиков потерь).

• **Критическим мышлением** для оценки целесообразности применения fine-tune в конкретном сценарии versus использования prompt engineering или RAG.

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

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

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

Владеет: Навыками работы с библиотеками и инструментами для обработки данных (Pandas, NumPy, Hugging Face Datasets).
Методами обеспечения репрезентативности и сбалансированности создаваемого набора данных.
Технологиями создания синтетических данных для задач, где реальных данных недостаточно.
Полным циклом подготовки данных: от сбора сырых данных до формирования готового для обучения объекта (DataLoader, Dataset)

MF-4 **Способен применять статистические методы для анализа данных, валидации моделей машинного обучения и проведения экспериментов в области ИИ.**

MF-4.1 Применяет статистические методы анализа и машинного обучения для решения задач анализа данных и проведения экспериментов на данных.
Применяет и выбирает методы статистического машинного обучения, учитывая особенности данных и задачи, а также объясняет различия между подходами.
Знает основные статистические методы описательного и инференционного анализа, принципы планирования экспериментов (A/B-тесты) и базовые алгоритмы машинного обучения (линейные модели, деревья).
Умеет применять статистические методы (проверка гипотез, анализ распределений) и алгоритмы машинного обучения для исследования данных, извлечения инсайтов и проверки рабочих гипотез.
Владеет навыком проведения полного цикла анализа данных: от предобработки и разведочного анализа (EDA) до построения, интерпретации результатов и формирования выводов.

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.
Строит модели динамических систем для многомерных временных рядов и полей.
Знает математические основы и предположения регрессионных, прогнозных моделей и методов анализа временных рядов (ARIMA, экспоненциальное сглаживание, подходы к работе с нестационарностью).
Умеет строить, обучать и валидировать предсказательные модели (регрессия, классификация, прогнозирование), включая работу с временными рядами и нестационарными процессами.
Владеет навыком выбора и настройки модели под конкретную задачу прогнозирования, диагностики её качества и интерпретации результатов прогноза.

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.
Оценивает статистические различия моделей и алгоритмов, обучаемых на данных.
Знает и применяет модифицированные статистические критерии, A/B тестирование. Применяет оценивание на основе модифицированных доверительных интервалов, использует Байесовские тесты.

MF-7 **Способен применять методы дифференциальной геометрии и топологии для формализации, анализа и интерпретации структур данных и признаковых пространств, включая задачи отображения, кластеризации, обучения на многообразиях и анализа устойчивости моделей.**

MF-7.1 Применяет методы топологического анализа для описания глобальных свойств данных и устойчивости признаковых структур.

Знает: Узнаёт и интерпретирует базовые топологические характеристики (связность, количество компонент, размерность) в примерах и визуализациях. **Умеет:** Использовать топологические дескрипторы в качестве новых признаков для модели классификации, характеризующих глобальную форму данных. Оценивать топологическую устойчивость признакового пространства к малым возмущениям в данных. **Владеет: Навыком** чтения и интуитивной интерпретации персистентных диаграмм для быстрой оценки сложности структуры данных.
• **Методом** использования TDA как инструмента для выявления неочевидных глобальных закономерностей, не улавливаемых традиционными статистическими методами.

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

ML-2.3 Решает проблемы несбалансированных данных и оценивает качество моделей

Знает проблемы несбалансированных данных методы оценивания качества моделей

Умеет применить на практике основные метрики оценки качества для задач классификации и регрессии (Б)

Применяет различные типы кросс-валидации Оценивает качество моделей с учетом bias-variance trade-off (П)

Владеет продвинутыми методами работы с несбалансированными данными (SMOTE weighted learning). Настраивает кастомные метрики и функции потерь. Проводит статистический анализ значимости 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 единицы (72 часа), их распределение по видам работ представлено в таблице

Виды работ	Всего часов	Форма обучения очная
		7 семестр (часы)
Контактная работа, в том числе:	**16,2**	**16,2**
Аудиторные занятия (всего):	**16,2**	**16,2**
занятия лекционного типа		
лабораторные занятия	16	16
практические занятия	-	-
семинарские занятия	-	-
Иная контактная работа:	**0,2**	**0,2**
Контроль самостоятельной работы (КСР)		
Промежуточная аттестация (ИКР)	0,2	0,2

Самостоятельная работа, в том числе:		**55,8**	**55,8**
Курсовая работа/проект (КР/КП) (подготовка)		-	-
Контрольная работа		-	-
Расчётно-графическая работа (РГР) (подготовка)		-	-
Выполнение индивидуальных заданий по подготовке рефератов, сообщений, презентаций		9,8	9,8
Самостоятельная проработка и материала учебников и учебных пособий, подготовка к лабораторным занятиям		38	38
Подготовка к текущему контролю		6	6
Контроль:			
Подготовка к экзамену		-	-
Общая трудоемкость	**час.**	**72**	**72**
	в том числе контактная работа	**16,2**	**16,2**
	зач. ед	**2**	**2**

2.2 Содержание дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.
Разделы/темы дисциплины, изучаемые в 7 семестре 4 курса очной формы обучения

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1.	Введение в аналитику данных для ML. От подготовки к анализу. Метрики качества классификации. Матрица ошибок, Accuracy, Precision, Recall, F1-score, ROC-AUC. Кросс-валидация.	8			2	6
2.	Планирование и проектирование признакового пространства (Feature Planning) Стратегии работы с дисбалансом классов.	10			2	8
3.	Продвинутый разведочный анализ (EDA) и диагностика проблем датасета.	10			2	8
4.	Анализ и управление целевой переменной - Стратегии работы с многоклассовой и мультилабельной классификацией.	8			2	6
5.	Стратегии балансировки и аугментации данных – методы семплирования, генеративные модели	8			2	6
6.	Анализ и оптимизация состава признакового пространства: Продвинутые методы селекции признаков	8			2	6
7.	Методология экспериментирования и статистическая оценка улучшений	10			2	8
8.	Проектирование сквозного пайплайна анализа данных Интерпретация моделей (SHAP, LIME).	9,8			2	7,8
	ИТОГО по разделам дисциплины	69,8			16	55,8
	Контроль самостоятельной работы (КСР)	-				
	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

Занятия лекционного типа не предусмотрены учебным планом.

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 аналитику данных для ML. От подготовки к анализу. Метрики качества классификации. Матрица ошибок, Accuracy, Precision, Recall, F1-score, ROC-AUC. Кросс-валидация.	Знакомство с библиотеками (pandas, numpy, matplotlib, scikit-learn). Формирование (поиск и составление) и первичный анализ набора данных. Разделение на train/test. Метрики качества классификации. Матрица ошибок, Accuracy, Precision, Recall, F1-score, ROC-AUC. Кросс-валидация.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
2.	Планирование и проектирование признакового пространства (Feature Planning) Стратегии работы с дисбалансом классов.	• Углублённый фичингениринг на основе доменных знаний и EDA. • Создание признаков взаимодействия, полиномиальных признаков. • Работа с временными рядами для создания признаков (лагги, скользящие статистики). • Анализ и привлечение внешних данных для обогащения датасета.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
3.	Продвинутый разведочный анализ (EDA) и диагностика проблем датасета.	• Анализ многомерных выбросов и их влияние на модель. • Методы обнаружения скрытых закономерностей и смещений (bias) в данных. • Статистические методы оценки стабильности датасета (Population Stability Index). • Диагностика "утечки данных" (data leakage) в признаках.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
4.	Анализ и управление целевой переменной - Стратегии работы с многоклассовой и мультилабельной классификацией.	• Углублённый анализ разметки. Методы поиска и исправления "шумных" меток. • Стратегии работы с многоклассовой и мультилабельной классификацией. • Постановка задач регрессии и анализ распределения целевой переменной.	Опрос по теоретическому материалу. Отчет по лабораторной работе.

5.	Стратегии балансировки и аугментации данных – методы семплирования, генеративные модели	• Сравнение и глубокий анализ методов сэмплирования (SMOTE, ADASYN, SMOTE-ENN). • Генеративные модели (VAE, GAN) для аугментации табличных данных (теория и практика). • Cost-sensitive learning как альтернатива сэмплингу.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
6.	Анализ и оптимизация состава признакового пространства: Продвинутые методы селекции признаков	• Продвинутые методы селекции признаков: взаимная информация, методы-обёртки (RFE, Boruta). • Анализ мультиколлинеарности и её влияние на различные модели. • Методы понижения размерности (PCA, t-SNE, UMAP) для анализа, но не для моделирования.	Опрос по теоретическому материалу. Отчет по лабораторной работе.
7.	Методология экспериментирования и статистическая оценка улучшений	• Протоколы A/B-тестирования различных версий датасета. • Статистические критерии для сравнения моделей (t-test, McNemar's test). • Анализ learning curves для диагностики недостатка данных или избыточности признаков. • Фреймворк для документирования экспериментов по улучшению данных (аналоги MLOps для данных).	Обсуждение вопросов по проекту
8.	Проектирование сквозного пайплайна анализа данных Интерпретация моделей (SHAP, LIME).	Защита проектов с использованием ML по заданию от индустриальных партнеров	Защиты проектов

2.3.4. Примерная тематика курсовых работ (проектов)

Курсовая работа не предусмотрена. В качестве курсового проекта студенты защищают инфраструктуру проекта приложения с использованием ML 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,	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры

	материала учебной и научной литературы, подготовка к семинарским занятиям	вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
2	Подготовка к лабораторным занятиям	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
3	Подготовка к решению задач и тестов	Методические указания по выполнению самостоятельной работы, утвержденные на заседании кафедры вычислительных технологий факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 07.05.2025 г.
4	Подготовка и выполнение проекта с использованием ML по заданию от индустриального партнера.	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: выполнение лабораторных работ 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:

- Технология использования компьютерных программ – позволяет эффективно дополнить процесс обучения языку на всех уровнях.
- Интернет-технологии – предоставляют широкие возможности для поиска информации, разработки научных проектов, ведения научных исследований.

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

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

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

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

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

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

Основные виды интерактивных образовательных технологий включают в себя:

- 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;
- проектная технология - индивидуальная или коллективная деятельность по отбору, распределению и систематизации материала по определенной теме, в результате которой составляется проект;
- анализ конкретных ситуаций - анализ реальных проблемных ситуаций, имевших место в соответствующей области профессиональной деятельности, и поиск вариантов лучших решений;
- 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 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 «Анализ данных машинного обучения».

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

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

освоения содержания дисциплины используются оценочные средства. Они направлены на определение степени сформированности компетенций.

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

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

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

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

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

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

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

4.1 Оценочные средства для текущего контроля успеваемости

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

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

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

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

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

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

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

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

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Контролируемые разделы (темы) дисциплины*	Код контролируемой компетенции (или ее части)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	Введение в аналитику данных для ML. От подготовки к анализу. Метрики качества классификации. Матрица ошибок, Accuracy, Precision, Recall, F1-score, ROC-AUC. Кросс-валидация.	BD-1, BD-2, MF-7, ML-2	*Лабораторная работа №1*	*Вопросы к зачету*
2	Планирование и проектирование признакового пространства (Feature Planning) Стратегии работы с дисбалансом классов.	BD-1, BD-2, MF-7, ML-2	*Лабораторная работа №2*	*Вопросы к зачету*
3	Продвинутый разведочный анализ (EDA) и диагностика проблем датасета.	BD-1, BD-2, LLM-2, MF-7, ML-2	*Лабораторная работа №3*	*Вопросы к зачету*
4	Анализ и управление целевой переменной - Стратегии работы с многоклассовой и мультилабельной классификацией.	BD-1, BD-2, LLM-2, MF-7, ML-2	*Лабораторная работа №4*	*Вопросы к зачету*
5	Стратегии балансировки и аугментации данных – методы семплирования, генеративные модели	BD-1, BD-2, LLM-2, MF-7, ML-2	*Лабораторная работа №5*	*Вопросы к зачету*
6	Анализ и оптимизация состава признакового пространства:	BD-1, BD-2, LLM-2, MF-7, ML-2	*Лабораторная работа №6*	*Вопросы к зачету*

	Продвинутые методы селекции признаков			
7	Методология экспериментирования и статистическая оценка улучшений	BD-1, BD-2, LLM-2, MF-7, ML-2	*Лабораторная работа №7*	*Вопросы к зачету*
8	Проектирование сквозного пайплайна анализа данных Интерпретация моделей (SHAP, LIME).	BD-1, BD-2, LLM-2, MF-7, ML-2	*Финальный проект*	*Вопросы к зачету*

Показатели, критерии и шкала оценки сформированных компетенций

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

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

BD-1.1 Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи
Знает методы заполнения пропусков в данных и удаления выбросов в табличных данных (случайные величины)
Имеет навыки (**умеет**) очистки зашумленных временных рядов и изображений. Обнаруживает и устраняет выбросы в данных временных рядов. **Владеет** подходами к заполнению пропусков в данных временных рядов и изображений.

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

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

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

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

BD-2.1 Знает, как сформировать требования для набора данных. Владеет умениями по формированию требований к наборам и качеству данных для решения задач машинного обучения

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

BD-2.3 Применяет инструменты и практики непрерывной интеграции данных (DataOps) **Умеет** применять инструменты интеграции данных. **Владеет** навыками непрерывной интеграции данных (DataOps)

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

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

Знает: основные подходы к тонкой настройке: полная настройка всех параметров, поэтапная разморозка слоев, методы эффективной тонкой настройки (P-Tuning, LoRA, QLoRA, Adapter). Гиперпараметры, критически важные для fine-tune: learning rate, scheduler, batch size, и их отличия от обучения с нуля.

Умеет: Отличать дообучение от первичного обучения, знает базовые процедуры **fine-tune,** анализировать задачу и выбирать наиболее подходящий метод fine-tune (полная настройка vs. эффективные методы).

Владеет: **Навыком** осознанного выбора стратегии fine-tune под ограничения (вычислительные ресурсы, объем данных, требования к качеству). Применяет fine-tune к предобученным моделям на новых датасетах.

• **Методами** анализа и интерпретации процесса дообучения (использование логов, графиков потерь).

• **Критическим мышлением** для оценки целесообразности применения fine-tune в конкретном сценарии versus использования prompt engineering или RAG.

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

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

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

Владеет: Навыками работы с библиотеками и инструментами для обработки данных (Pandas, NumPy, Hugging Face Datasets). **Методами** обеспечения репрезентативности и сбалансированности создаваемого набора данных.

Технологиями создания синтетических данных для задач, где реальных данных недостаточно.

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

LLM-2.3 **Использует адаптивные методы дообучения**

Знает адаптивные методы дообучения.

Умеет применять базовые адаптивные методы (prefix, adapter). Выбирать методы с учетом требований к latency и ресурсам.

MF-7 Способен применять методы дифференциальной геометрии и топологии для формализации, анализа и интерпретации структур данных и признаковых пространств, включая задачи отображения, кластеризации, обучения на многообразиях и анализа устойчивости моделей.

MF-7.1 Применяет методы топологического анализа для описания глобальных свойств данных и устойчивости признаковых структур.

Знает: Узнаёт и интерпретирует базовые топологические характеристики (связность, количество компонент, размерность) в примерах и визуализациях. **Умеет:** Использовать топологические дескрипторы в качестве новых признаков для модели классификации, характеризующих глобальную форму данных. Оценивать топологическую устойчивость признакового пространства к малым возмущениям в данных. **Владеет: Навыком** чтения и интуитивной интерпретации персистентных диаграмм для быстрой оценки сложности структуры данных.

• **Методом** использования TDA как инструмента для выявления неочевидных глобальных закономерностей, не улавливаемых традиционными статистическими методами.

MF-7.2 Использует геометрические представления данных (например, многообразия) при построении или интерпретации моделей машинного обучения

Знает: Принципы работы методов, явно использующих геометрию данных: Manifold Learning (Isomap, LLE, t-SNE, UMAP). Понимает, что данные могут лежать на многообразии; использует методы визуализации (PCA, t-SNE, UMAP) для исследования структуры данных. Умеет: Оценивать внутреннюю размерность данных. Применять алгоритмы обучения на многообразии (Manifold Learning) для нелинейного снижения размерности перед построением модели классификации. Владеет: Навыками выбора между линейными (PCA) и нелинейными (UMAP, t-SNE) методами снижения размерности в зависимости от геометрии данных.

MF-7.3 Обосновывает выбор методов визуализации или трансформации признаков с учётом их топологических и метрических свойств.

Знает: Цели и ограничения различных методов визуализации (PCA, t-SNE, UMAP, MDS) с точки зрения сохранения топологии и метрики. Какие свойства сохраняют линейные (PCA) и нелинейные (t-SNE, UMAP, Isomap) методы (глобальная структура, локальные окрестности, топологические инварианты).

Умеет: выбирать методы визуализации и трансформации признаков из предложенного набора; знает, что разные методы отображают разные аспекты структуры данных.

Владеет: Навыком построения и критического анализа визуализаций высокоразмерных данных, понимания их достоинств и недостатков. Методологией последовательного применения методов: от топологического анализа (ИД-1) для понимания структуры к выбору адекватного метода визуализации/трансформации

(ИД-3). Способностью обосновать стейкхолдерам выбор того или иного подхода к анализу и визуализации данных для задачи классификации.

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)

ML-2.2 Применяет методы предварительной обработки данных и работы с признаками

Знает методы предварительной обработки данных и работы с признаками

Умеет проводить разведочный анализ данных, умеет работать с пропущенными значениями и выбросами (Б). **Умеет** проектировать и внедрять комплексные пайплайны предварительной обработки данных с использованием современных методов ИИ, автоматизации и feature engineering в различных предметных областях (Э)

Владеет методами feature engineering: отбор создание и преобразование признаков. (П)

ML-2.3 Решает проблемы несбалансированных данных и оценивает качество моделей

Знает проблемы несбалансированных данных методы оценивания качества моделей

Умеет применить на практике основные метрики оценки качества для задач классификации и регрессии (Б)

Применяет различные типы кросс-валидации Оценивает качество моделей с учетом bias-variance trade-off (П)

Владеет продвинутыми методами работы с несбалансированными данными (SMOTE weighted learning). Настраивает кастомные метрики и функции потерь. Проводит статистический анализ значимости 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

Сквозные требования ко всем лабораторным работам:

Прототип: Код на Python или R (Jupyter Notebook или скрипты).

Отчет:

Требуемые разделы отчета:

- **Постановка задачи: Описание задачи, EDA,**
- **Теоретическая справка** - краткое описание подходов решения (1-2 абзаца), предобработка данных, выбранные модели, метрики качества, выводы.

- **Экспериментальная часть** - код и результаты выполнения всех частей работы.
- **Визуализации** – графики, необходимые для лучшего понимания и восприятия результатов

Презентация: Краткий разбор решения для "защиты" перед условным индустриальным партнером.
Примечание: При разработке лабораторных были использованы кейсы от индустриальных партнеров, перечисленные ниже в разделе 6.

Примеры лабораторных работ

Лабораторная работа №1: Текстовый анализ и классификация пользовательских обращений

Тема: Основы NLP: предобработка текста, векторизация, классификация, анализ тональности.
Связь с программой: Введение в ML, линейные модели, логистическая регрессия, методы векторизации текста (TF-IDF, Word2Vec).
Индустриальные кейсы:

№2: НЛП-анализ жалоб клиентов в свободной форме (Основной кейс)
№5: Объяснимость и контроль генеративных моделей (Элемент объяснимости)

Задачи для студентов:

1. Получить датасет с синтетическими или реальными (анонимизированными) текстами обращений клиентов и их метками (категория проблемы, тональность).
2. Провести EDA: распределение тем, длина текстов, частотность слов.
3. Реализовать пайплайн предобработки: лемматизация, удаление стоп-слов, очистка.
4. Обучить несколько моделей (например, Logistic Regression, Random Forest, Naive Bayes) на TF-IDF векторизованных текстах для multi-label классификации (тема + тональность).
5. Оценить модели по метрикам precision, recall, F1-score.
6. **Задание повышенной сложности:** Использовать предобученные эмбеддинги (FastText, ruBERT) и сравнить результаты. Применить SHAP/LIME для объяснения предсказаний модели по конкретным примерам.

Ожидаемый результат: Прототип, который по тексту обращения предсказывает его тему и эмоциональную окраску, с обоснованием выбора модели и ее интерпретацией.

Лабораторная работа №2: Генеративные языковые модели (LLM) для анализа и создания контента

Тема: Применение LLM для задач генерации и суммаризации текста. Fine-tuning, Prompt Engineering, RAG.

Связь с программой: Архитектуры нейросетей, трансформеры, transfer learning.
Индустриальные кейсы:

№1: Генеративный ИИ для инвестиционных обзоров (Основной кейс)
№3: Генерация сценариев фишинговых писем (Дополнительный кейс)
№6: Генерация пользовательских сценариев (Дополнительный кейс)

Задачи для студентов:

1. **Кейс №1:** На основе структурированных данных (таблица с котировками, макро-показателями) и нескольких примеров готовых обзоров разработать систему генерации отчетов.
2. Использовать Prompt Engineering для большой LLM (например, через API: GPT-4, YandexGPT, SberAI) для создания текстов по шаблону.
3. **Задание повышенной сложности:** Реализовать простейшую RAG-систему: загрузить в векторную базу данных исторические отчеты и использовать их как контекст для генерации более качественных и релевантных текстов.
4. **Кейс №3 (на выбор):** Использовать ту же LLM для генерации фишинговых писем по заданным параметрам (тема, стиль). Провести анализ "реалистичности" сгенерированных примеров.

Ожидаемый результат: Инструмент, принимающий на вход таблицу с данными и возвращающий текстовый инвестиционный обзор в заданном формате. А также набор сгенерированных фишинговых сценариев.

Лабораторная работа №3: Создание и валидация синтетических данных

Тема: Генеративные модели (GAN, VAEs, Diffusion) для создания tabular и sequential data. Метрики оценки синтетических данных.
Связь с программой: Генеративные модели, оценка качества моделей, статистические тесты.
Индустриальные кейсы:

№7: Генерация synthetic data для банковских моделей (Основной кейс)
№9: Оценка кредитоспособности на основе поведенческих данных (Источник данных для следующей работы)

Задачи для студентов:

1. Получить реальный (или близкий к реальному) анонимизированный датасет с транзакциями или клиентскими профилями.
2. Обучить модель для генерации синтетических данных (например, CTGAN, TVAE).
3. Сгенерировать синтетический датасет сопоставимого размера.
4. Провести валидацию:
   - **Статистические тесты:** Сравнение распределений отдельных признаков (K-S test).
   - **Визуализация:** Построение t-SNE/UMAP карт для визуального сравнения реальных и синтетических точек.
   - **Метрики:** Рассчет расстояния между распределениями (например, Wasserstein distance).
5. **Задание повышенной сложности:** Проверить "полезность" синтетических данных, обучив на них простую ML-модель (например, для предсказания дефолта) и проверив ее качество на реальном тестовом наборе.

Ожидаемый результат: Синтетический датасет и отчет с детальным сравнением его с исходными данными, включая визуализации и расчет метрик.

Лабораторная работа №4: Построение и интерпретация прогнозных ML-моделей для принятия решений

Тема: Регрессия, классификация, feature engineering, explainable AI (XAI).
Связь с программой: Деревья решений, ансамбли (Random Forest, Gradient Boosting), методы интерпретации моделей.
Индустриальные кейсы:

№8: Модель анализа инвестиционной привлекательности МСП (Основной кейс)
№9: Индивидуальная оценка кредитоспособности (Основной кейс)
№10: Предиктивная аналитика ROI (Дополнительный кейс)

Задачи для студентов (на выбор один из двух основных кейсов):

1. **Кейс №8 или №9:** Получить датасет с признаками компаний МСП или клиентов (финансовые, поведенческие).
2. Провести детальный Feature Engineering: создание новых признаков, отбор наиболее важных.
3. Обучить модель градиентного бустинга (CatBoost, LightGBM, XGBoost) для предсказания целевой переменной (рейтинг/вероятность дефолта).
4. Оценить модель на hold-out выборке, используя релевантные метрики (Accuracy, F1, ROC-AUC).
5. **Обязательный элемент:** Применить методы XAI (SHAP, LIME) для анализа модели.
   - Построить график важности признаков (global explanation).
   - Дать интерпретацию прогноза для нескольких конкретных объектов (local explanation). Например: "Заявке клиента X присвоен низкий рейтинг из-за высокого соотношения долга к доходу и низкой активности в приложении".

Ожидаемый результат: Обученная модель с высокой предсказательной силой и подробный отчет, объясняющий, какие факторы больше всего влияют на прогноз и почему модель приняла то или иное решение для конкретных кейсов.

Лабораторная работа №5: Сравнительный анализ мультимодальных и генеративных моделей (Project Capstone)

Тема: Работа в команде, планирование экспериментов, сравнение сложных моделей, мультимодальность.
Связь с программой: Все ранее изученные темы, командная работа, презентация результатов.
Индустриальные кейсы:

- **№12: Сравнение text2video / text2img моделей** (Основной кейс)
- **№4: Мультимодальный ассистент** (Элемент проектирования)
- **№11: Анализ поведения в экосистеме цифрового рубля** (Альтернативный кейс для анализа данных)

Задачи для студентов (командный проект 3-4 человека):

1. **Основной путь (№12):**
   - Выбрать 3-4 открытые модели text2image (Stable Diffusion, DALL-E, Midjourney через API) и/или text2video (ModelScope, Sora API если доступен).
   - Разработать единый набор промптов (запросов) разной сложности (простой объект, сложная сцена, стиль).
   - Запустить генерацию для всех моделей на одном и том же наборе промптов, фиксируя ресурсы (время, VRAM).
   - Разработать схему оценки: объективные метрики (например, CLIP-score) и субъективные (human evaluation: голосование по качеству, релевантности, артефактам).
   - Подготовить финальный отчет-таблицу с результатами и выводами.
2. **Альтернативный путь (№4):** Разработать концептуальный прототип мультимодального ассистента, спроектировав его архитектуру (модуль распознавания

речи -> NLP-движок -> модуль синтеза речи + модуль компьютерного зрения) и описав, какие готовые модели (Whisper, GPT, YOLO) можно использовать для каждого компонента.

Ожидаемый результат: Для №12 - детализированный отчет-исследование с примерами генераций и выводами о применимости моделей. Для №4 - архитектурный дизайн-документ и демо-прототип на стыке 2-3 модальностей (например, речь -> текст -> ответ).

Критерии оценки работ:

Качество кода (20 баллов) - читаемость, комментарии, эффективность

Полнота выполнения (30 баллов) - все части работы выполнены

Визуализации (20 баллов) - качество и информативность графиков

Анализ и выводы (30 баллов) - глубина анализа, практические рекомендации

Максимальный балл: 100

Зачетно-экзаменационные материалы для промежуточной аттестации (зачет)

Общие принципы:

1. Что такое переобучение (overfitting) и недообучение (underfitting)? Как их диагностировать?
2. Что такое кросс-валидация и зачем она нужна? Опишите метод k-fold cross-validation.
3. Почему простое разбиение на обучающую и тестовую выборку может быть недостаточным? Что такое переобучение (overfitting) и как разбиение данных помогает с ним бороться?
4. Что такое стратифицированное разбиение (stratified split)? В каких случаях его необходимо использовать и почему?
5. Какие методы кодирования категориальных признаков вы знаете? В чем разница между One-Hot Encoding и Label Encoding? Когда какой метод предпочтительнее?

Основы дифференциальной геометрии и топологии в интеллектуальном анализе данных

6. Объясните "гипотезу многообразия" (Manifold Hypothesis) простыми словами. Приведите пример из реальной жизни.
7. Что такое intrinsic (внутренняя) размерность данных? Почему она важна и как она связана с "проклятием размерности"?
8. Какие базовые топологические инварианты можно использовать для описания структуры данных? Что такое связность, и как она связана с кластеризацией?
9. Как методы снижения размерности, такие как PCA, t-SNE и UMAP, связаны с геометрией и топологией данных? Какой из них лучше сохраняет глобальную структуру, а какой — локальную?
10. Для чего используется алгоритм UMAP и в чем его основное топологическое преимущество перед t-SNE?

Кросс-валидация

11. Объясните принцип работы k-кратной кросс-валидации (k-fold cross-validation). Каковы ее преимущества перед простым hold-out разбиением?
12. Что такое стратифицированная k-кратная кросс-валидация? Почему она важна для несбалансированных данных?
13. В чем разница между параметрами модели и гиперпараметрами? Как кросс-валидация используется для подбора гиперпараметров?

Стратегии работы с дисбалансом классов

18. Какие существуют три основных подхода к борьбе с дисбалансом классов (на уровне данных, на уровне алгоритма и ансамблирование)?
19. В чем разница между Random Undersampling и Random Oversampling? Каковы главные недостатки каждого из этих "наивных" методов?
20. Опишите принцип работы алгоритма SMOTE для генерации синтетических примеров минорного класса. Какие у него есть преимущества перед простым дублированием?
21. Как работает метод взвешивания классов (class weighting)? Где именно в процессе обучения модели он применяется?
22. Какую метрику следует выбрать в качестве основной для оптимизации при работе с сильно несбалансированными данными и почему?

Feature Engineering

23. Что такое Feature Engineering и почему это один из самых важных этапов в построении ML-модели?
24. Какие вы знаете методы создания новых признаков? Приведите примеры для табличных, текстовых и временных данных.
25. Зачем нужно масштабировать и нормализовать признаки? Какие алгоритмы критически зависят от этой процедуры (приведите примеры), а какие — нет?
26. Объясните разницу между отбором признаков (feature selection) и выделением признаков (feature extraction). Какие методы относятся к каждому из этих подходов?
27. Что такое "проклятие размерности" и как Feature Engineering помогает с ним бороться?

Блок 1: Фундаментальные концепции и индустриальный контекст

28. Опишите жизненный цикл ML-проекта в индустрии. Какие этапы являются наиболее критичными для успешного внедрения в такой регулируемой сфере, как банковское дело?
29. В чем заключаются основные проблемы и ограничения при работе с реальными данными в финансовой отрасли (на примере задач из курса)? Как вы с ними боролись?
30. Объясните разницу между подходами Proof-of-Concept (POC), Prototype и Production-ready моделью. На каком этапе находятся решения, которые вы разрабатывали в ходе лабораторных работ?

Блок 2: Обработка естественного языка (NLP) и анализ текста

31. Опишите полный пайплайн предобработки текста для задачи классификации жалоб клиентов. Почему такие шаги, как лемматизация и удаление стоп-слов, важны?
32. Сравните подходы к векторизации текста: TF-IDF и Word Embeddings (например, Word2Vec, FastText). В каких сценариях предпочтительнее каждый из них, исходя из вашего опыта в ЛР №1?
33. Какие метрики качества наиболее уместны для оценки многоклассовой классификации с несбалансированными данными (как в задаче с жалобами)? Обоснуйте свой выбор.

Блок 3: Генеративные модели и Large Language Models (LLM)

34. В чем разница между подходами Fine-tuning и Prompt Engineering при работе с большими языковыми моделями? Приведите примеры из ЛР №2, где уместен каждый из подходов.
35. Что такое RAG (Retrieval-Augmented Generation) и как эта архитектура помогает решить проблему «галлюцинаций» LLM в таких задачах, как генерация инвестиционных обзоров?

36. Какие этические и безопасностные риски возникают при использовании генеративных моделей (например, для создания фишинговых писем или синтетических данных)? Как их можно минимизировать?

Блок 4: Генерация и валидация синтетических данных

37. Сформулируйте основные требования к качеству синтетических данных для обучения банковских моделей. Почему недостаточно просто скопировать распределения отдельных признаков?
38. Опишите методы и метрики, которые вы использовали для валидации синтетических данных в ЛР №3. Что такое «полезность» (utility) синтетических данных и как ее измерить?
39. Каковы правовые и регуляторные аспекты использования синтетических данных в контексте Федерального закона № 152-ФЗ «О персональных данных»?

Блок 5: Интерпретируемость и объяснимость моделей (XAI)

40. Почему Explainable AI (XAI) является критически важным требованием для ML-моделей в банковской сфере? Приведите примеры из задач скоринга или анализа привлекательности бизнеса.
41. Различайте глобальную и локальную интерпретируемость моделей. Какие методы (например, SHAP, LIME) для чего лучше подходят?
42. Как бы вы объяснили руководителю без технического бэкграунда, почему модель отказала в кредите конкретному заемщику, используя результаты SHAP-анализа?

Блок 6: Сравнительный анализ и мультимодальность

43. Опишите методику сравнительного анализа сложных генеративных моделей (text2img/text2video). Какие объективные и субъективные метрики вы считаете наиболее показательными и почему?
44. Каковы основные архитектурные 挑战 (challenges) при создании мультимодального ассистента, объединяющего речь, текст и компьютерное зрение?
45. Представьте, что вы получили задачу проанализировать паттерны поведения пользователей цифрового рубля (Кейс №11). Опишите ваш план действий: с какими данными будете работать, какие гипотезы проверите, какие модели примените.

Блок 7: Интеграция, внедрение и будущее

46. Что такое MLOps и как его принципы могут быть применены к одному из проектов, которые вы выполняли в ходе курса, для перевода прототипа в промышленную эксплуатацию?
47. Проанализируйте тренды в области ИИ, которые, на ваш взгляд, окажут наибольшее влияние на финтех-индустрию в ближайшие 3-5 лет. Обоснуйте свой выбор, опираясь на опыт, полученный в ходе курса.

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 «Анализ данных машинного обучения» 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.

4.3. Методические указания по организации вычислительной инфраструктуры

Требования к аппаратному и программному обеспечению рабочих мест

Аппаратные требования.

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

Программные требования.

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

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

Инструменты и библиотеки:

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

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

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

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

5.1.Основная литература

(в том числе публикации конференций A*)

1. 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
2. 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
4. Kadurin, Artur, et al. "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology." Oncotarget 8.7 (2016): 10883.
5. 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.
6. Polykovskiy, Daniil, et al. "Molecular sets (MOSES): a benchmarking platform for molecular generation models." Frontiers in pharmacology 11 (2020): 565644.
7. 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.
8. Polykovskiy, Daniil, et al. "Entangled conditional adversarial autoencoder for de novo drug discovery." Molecular pharmaceutics 15.10 (2018): 4398-4405.
9. 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.

10. 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.
11. 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
12. Ярушкина, Н. Г. Интеллектуальный анализ временных рядов : учебное пособие для студентов вузов / Ярушкина, Надежда Глебовна, Т. В. Афанасьева, И. Г. Перфильева ; Н. Г. Ярушкина, Т. В. Афанасьева, И. Г. Перфильева. - М. : ФОРУМ : ИНФРА-М, 2012. - 159 с. : ил. - (Высшее образование). - Библиогр. в конце глав. - ISBN 9785819904961. - ISBN 9785160051970.

5.2. Дополнительная литература:

1. Подоплелова, Е. С. Современные методы инженерии знаний в задачах машинного обучения : учебное пособие : [16+] / Е. С. Подоплелова ; Южный федеральный университет, Инженерно-технологическая академия. – Ростов-на-Дону ; Таганрог : Южный федеральный университет, 2025. – 130 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=724468 (дата обращения: 30.10.2025). – Библиогр. в кн. – ISBN 978-5-9275-4882-8. – Текст : электронный.
2. Протодьяконов, А. В. Алгоритмы Data Science и их практическая реализация на Python : учебное пособие : [16+] / А. В. Протодьяконов, П. А. Пылов, В. Е. Садовников. – Москва ; Вологда : Инфра-Инженерия, 2022. – 392 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=725623 (дата обращения: 30.10.2025). – Библиогр.: с. 380-383. – ISBN 978-5-9729-1006-9. – Текст : электронный.
3. Пылов, П. А. Основы работы с моделями машинного и глубокого обучения : учебное пособие : [16+] / П. А. Пылов, Р. В. Майтак, А. В. Дягилева. – Москва ; Вологда : Инфра-Инженерия, 2023. – 256 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=725673 (дата обращения: 30.10.2025). – Библиогр.: с. 247-250. – ISBN 978-5-9729-1547-7. – Текст : электронный.
4. Татарникова, Т. М. Интеллектуальный анализ данных : учебное пособие : [16+] / Т. М. Татарникова. – Москва ; Вологда : Инфра-Инженерия, 2024. – 172 с. : ил., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=725643 (дата обращения: 30.10.2025). – Библиогр. в кн. – ISBN 978-5-9729-1772-3. – Текст : электронный.

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

Конференции A*:

1. https://openreview.net/forum?id=FMMF1a9ifL
2. https://openreview.net/forum?id=ElUrNM9U8c№discussion
3. https://openreview.net/forum?id=JoO6mtCLHD
4. https://aclanthology.org/2024.findings-emnlp.760/
5. https://aclanthology.org/2020.coling-main.588/
6. https://link.springer.com/chapter/10.1007/978-3-030-72113-8_30
7. https://link.springer.com/chapter/10.1007/978-3-031-42448-9_10
8. https://aclanthology.org/2024.findings-naacl.288/

Электронно-библиотечные системы (ЭБС):

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

Профессиональные базы данных

1. Scopus http://www.scopus.com/
2. ScienceDirect https://www.sciencedirect.com/
3. Журналы издательства Wiley https://onlinelibrary.wiley.com/
4. Научная электронная библиотека (НЭБ) http://www.elibrary.ru/
5. Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН http://archive.neicon.ru
6. Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/
7. Президентская библиотека им. Б.Н. Ельцина https://www.prlib.ru/
8. База данных CSD Кембриджского центра кристаллографических данных (CCDC) https://www.ccdc.cam.ac.uk/structures/
9. Springer Journals: https://link.springer.com/
10. Springer Journals Archive: https://link.springer.com/
11. Nature Journals: https://www.nature.com/
12. Springer Nature Protocols and Methods: https://experiments.springernature.com/sources/springer-protocols
13. Springer Materials: http://materials.springer.com/
14. Nano Database: https://nano.nature.com/
15. Springer eBooks (i.e. 2020 eBook collections): https://link.springer.com/
16. "Лекториум ТВ" http://www.lektorium.tv/
17. Университетская информационная система РОССИЯ http://uisrussia.msu.ru

Информационные справочные системы

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

Ресурсы свободного доступа

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

273--84d1f.xn--p1ai/voprosy_i_otvety.

Собственные электронные образовательные и информационные ресурсы КубГУ

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

6. Методические указания для обучающихся по освоению дисциплины

6.1 Рекомендации по организации обучения

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

Лабораторные занятия курса посвящены практическому освоению методов классификации. На занятиях студенты реализуют задачи классификации различных 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

1. **Подготовка данных**:

– Освойте методы обработки данных (нормализация, кодирование категориальных признаков) на примере датасетов Iris, MNIST/CIFAR-10.

– Используйте pandas для анализа и scikit-learn для разделения выборки (train/test split).

2. **Эксперименты с моделями**:

– Начинайте с простых алгоритмов (логистическая регрессия), постепенно переходя к сложным (ансамбли, нейросети).

– Сравнивайте результаты по метрикам (F1-score, ROC-AUC) для разных моделей.

3. **Анализ результатов**:

– Визуализируйте матрицы ошибок, кривые обучения.

– Интерпретируйте работу моделей с помощью SHAP/LIME (Lab 9).

– Оптимизируйте гиперпараметры через GridSearchCV.

6.3 Проектная деятельность

Итоговая работа — решение сквозной задачи классификации (например, прогнозирование оттока клиентов, распознавание объектов на изображениях). Этапы:

1. Выбор датасета и постановка задачи.
2. Предобработка данных, feature engineering.
3. Обучение и валидация моделей (минимум 3 разных алгоритма).
4. Интерпретация результатов и формирование отчета.

Рекомендуется использовать Kaggle для поиска datasets и соревнований.

6.4 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.

Кейс от 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
3. 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
2	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
3	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
4	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.