

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

Б1.В.ДВ.02.02
(индекс дисциплины)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Интерпретируемый искусственный интеллект: методы и анализ решений моделей

(наименование дисциплины)

по направлению подготовки
09.03.04 Программная инженерия

направленность (профиль)
Программная инженерия с применением ИИ-технологий

Форма обучения: заочная

Год набора: 2025

Общая трудоемкость: 5 ЗЕ

Распределение часов дисциплины по семестрам

Семестр	7	Итого
Форма контроля	Зачет	
Вид занятий		
Лекции	6	6
Лабораторные		
Практические		
Руководство: курсовые работы (проекты) / РГР		
Промежуточная аттестация	0,25	0,25
Контактная работа	6,25	6,25
Самостоятельная работа	170	170
Контроль	3,75	3,75
Итого	180	180

Рабочую программу составил:

доцент института цифровых технологий, канд.техн.наук, доцент Аникина О.В.

(должность, ученое звание, степень, Фамилия И.О.)

Рецензирование рабочей программы дисциплины:



Отсутствует



Рецензент

(должность, ученое звание, степень, Фамилия И.О.)

Рабочая программа дисциплины составлена на основании ФГОС ВО и учебного плана направления подготовки

09.03.04 Программная инженерия

Срок действия рабочей программы дисциплины до «31» августа 2031 г.

УТВЕРЖДЕНО

На заседании института цифровых технологий

(протокол заседания № 1 от «05» сентября 2025 г.).

1. Цель освоения дисциплины

Цель освоения дисциплины – формирование у обучающихся умения применять методы интерпретируемого искусственного интеллекта (ХАИ) для анализа, объяснения и верификации решений моделей машинного обучения.

2. Место дисциплины в структуре ОПОП ВО

Дисциплины и практики, на освоении которых базируется данная дисциплина: «Управление требованиями к программному обеспечению».

Дисциплины и практики, для которых освоение данной дисциплины необходимо как предшествующее: «Тестирование программного обеспечения», «Инженерия машинного обучения для производственных процессов».

3. Планируемые результаты обучения

Формируемые и контролируемые компетенции (код и наименование)	Индикаторы достижения компетенций (код и наименование)	Планируемые результаты обучения
ПК -3. Способен проектировать тестовые сценарии и проводить проверку работоспособности программного обеспечения	ПК-3.1. Знает виды тестирования программного кода	Знать: уровни тестирования (модульное, интеграционное, системное); виды тестирования (функциональное, нагрузочное, регрессионное, приемочное). Уметь: выбирать виды тестирования для разных этапов проекта. Владеть: терминологией в области тестирования ПО.
	ПК-3.2. Умеет выполнять проверку работоспособности программного обеспечения	Знать: критерии качества ПО и приемочные критерии. Уметь: разрабатывать тестовые случаи (test cases) и тестовые сценарии. Владеть: навыками ручного тестирования и работы с баг-трекингowymi системами.
	ПК-3.3. Владеет навыками создания тестовых сценариев и использования их для проверки работоспособности программного обеспечения	Знать: техники тест-дизайна (классы эквивалентности, анализ граничных значений). Уметь: автоматизировать тестовые сценарии. Владеть: навыками работы с фреймворками для автоматизированного тестирования.

4. Структура и содержание дисциплины

Модуль (раздел)	Вид учебной работы	Наименование тем занятий (учебной работы)	Семестр	Объем, ч.	Баллы	Интерактив, ч.	Формы текущего контроля (наименование оценочного средства)
Модуль1. Фундаментальные основы и модели интерпретации	Лек. 1	Тема 1. Проблема «черного ящика» и базовые понятия ХАИ.	7	2	-	-	Отчеты по практическим работам 1-4
	Ср	Тема 2. Классификация методов интерпретации.	7	6	-	-	
	Ср	Тема 3. Интерпретация простых моделей: линейные модели и деревья решений.	7	6	-	-	
	Ср	Тема 4. Глобальная интерпретация: важность и влияние признаков.	7	6	-	-	
	Ср	Тема 5. Локальная интерпретация: теория и алгоритм LIME.	7	6	-	-	
	Ср	Тема 6. Локальная интерпретация: теория и алгоритм SHAP.	7	6	-	-	
	Ср	Тема 7. Построение и применение суррогатных моделей.	7	6	-	-	
	Лек. 2	Тема 8. Метрики для оценки качества интерпретаций.	7	2	-	-	
	Ср	ПР1. Анализ встроенной интерпретируемости линейных моделей и деревьев (часть 1)	7	6	-	-	
	Ср	ПР1. Анализ встроенной интерпретируемости линейных моделей и деревьев (часть 2)	7	6	7	-	

Модуль (раздел)	Вид учебной работы	Наименование тем занятий (учебной работы)	Семестр	Объем, ч.	Баллы	Интерактив, ч.	Формы текущего контроля (наименование оценочного средства)
	Ср	ПР2. Расчет и визуализация глобальной важности признаков (часть 1)	7	6	-	-	
	Ср	ПР2. Расчет и визуализация глобальной важности признаков (часть 2)	7	6	7	-	
	Ср	ПР3. Применение LIME для объяснения отдельных прогнозов. (часть 1)	7	6	-	-	
	Ср	ПР3. Применение LIME для объяснения отдельных прогнозов (часть 2)	7	6	7	-	
	Ср	ПР4. Применение SHAP для анализа моделей и прогнозов (часть 1)	7	6	-	-	
	Ср	ПР4. Применение SHAP для анализа моделей и прогнозов (часть 2)	7	6	7	-	
Модуль2. Интерпретация глубоких моделей и интеграция в процессы разработки ПО	Лек. 3	Тема 9. Особенности интерпретации сверточных нейронных сетей (CNN)	7	2	-	-	Отчеты по практическим работам 5-8
	Ср	Тема 10. Методы визуализации активаций и карт внимания в CNN	7	6	-	-	
	Ср	Тема 11. Attention mechanisms как инструмент интерпретации в NLP	7	6	-	-	
	Ср	Тема 12. Интерпретация моделей для обработки текстовых данных	7	6	-	-	
	Ср	Тема 13. Визуализация результатов XAI для различных стейкхолдеров	7	6	-	-	
	Ср	Тема 14. Интеграция XAI в цикл разработки и тестирования ПО (MLOps)	7	6	-	-	

Модуль (раздел)	Вид учебной работы	Наименование тем занятий (учебной работы)	Семестр	Объем, ч.	Баллы	Интерактив, ч.	Формы текущего контроля (наименование оценочного средства)
	Ср	Тема 15. Разработка тестовых сценариев для проверки корректности моделей	7	6	-	-	
	Ср	Тема 16. Этические аспекты и карьера в области интерпретируемого ИИ	7	6	-	-	
	Ср	ПР5. Визуализация карт активации и атрибуции в нейронных сетях (часть 1)	7	6	-	-	
	Ср	ПР5. Визуализация карт активации и атрибуции в нейронных сетях (часть 2)	7	6	7	-	
	Ср	ПР6. Анализ внимания в моделях для обработки текста (часть 1)	7	6	-	-	
	Ср	ПР6. Анализ внимания в моделях для обработки текста (часть 2)	7	6	7	-	
	Ср	ПР7. Сравнительный анализ методов ХАІ на комплексном кейсе (часть 1)	7	6	-	-	
	Ср	ПР7. Сравнительный анализ методов ХАІ на комплексном кейсе (часть 2)	7	6	9	-	
	Ср	ПР8. Разработка тестового сценария для верификации объяснений модели	7	8	9	-	
	ПА	Промежуточная аттестация	7	0,25		-	
	Контроль	Зачет	7	3,75	40	-	Итоговый тест
Итого:				180			

Схема расчета итогового балла: по накопительному рейтингу
Тебущий рейтинг + Результат итогового теста .

65. Образовательные технологии

В рамках учебного курса предусмотрены следующие образовательные технологии:

- Технологии традиционного обучения в форме лекций, практических работ и самостоятельной работы обучающихся.
- Технология проектного обучения: реализация и защита отчетов по практическим работам, выполнение и защита курсовой работы.

Технологии традиционного обучения - организация учебного процесса в вузе, основанная на лекционных и практических формах обучения: объяснительно-иллюстративное обучение. Данная технология применяется во всех модулях курса.

Технология интерактивного обучения - организация учебного процесса, которая предполагает максимальную активность обучающихся в процессе формирования ключевых компетенций. На практическом занятии обучающиеся представляют результат выполнения заданной работы.

6. Методические указания по освоению дисциплины

6.1. Рекомендации по подготовке к практическим занятиям

Обучающимся следует:

- до очередного практического занятия по рекомендованным литературным источникам проработать теоретический материал, соответствующей темы занятия;
- при подготовке к практическим занятиям следует обязательно использовать не только лекции, учебную литературу, но и другие источники;
- в начале занятий задать преподавателю вопросы по материалу, вызвавшему затруднения в его понимании и освоении при решении задач, заданных для самостоятельного решения;
- на занятии доводить каждую задачу до окончательного решения, демонстрировать понимание проведенных расчетов (анализов, ситуаций), в случае затруднений обращаться к преподавателю.

Для того чтобы практические занятия приносили максимальную пользу, необходимо помнить, что упражнение и решение задач проводятся по рассмотренному на лекциях материалу и связаны, как правило, с детальным разбором отдельных вопросов лекционного курса. Следует подчеркнуть, что только после усвоения лекционного материала с определенной точки зрения (а именно с той, с которой он излагается на лекциях) он будет закрепляться обучающимся на практических занятиях как в результате обсуждения и анализа лекционного материала, так и с помощью решения проблемных ситуаций, задач. При этих условиях обучающийся не только хорошо усвоит материал, но и научится применять его на практике, а также получит дополнительный стимул (и это очень важно) для активной проработки лекции.

При самостоятельном решении задач нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если обучающихся видит несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи). Решение проблемных задач или примеров следует излагать подробно, вычисления располагать в строгом порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что решение каждой учебной задачи должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом. Полученный ответ следует проверить способами, вытекающими из существа данной задачи. Полезно также (если возможно) решать несколькими способами и сравнить полученные результаты. Решение задач данного типа нужно продолжать до приобретения твердых навыков в их решении.

6.2. Рекомендации по подготовке к зачету

Подготовка к зачету способствует закреплению, углублению и обобщению знаний, получаемых, в процессе обучения, а также применению их к решению практических задач. Готовясь к зачету, обучающийся ликвидирует имеющиеся пробелы в знаниях, углубляет, систематизирует и упорядочивает свои знания. На зачете обучающийся демонстрирует то, что он приобрел в процессе обучения по учебной дисциплине.

Необходимо ориентировать обучающихся на систематическую подготовку к занятиям в течение семестра, что позволит использовать время зачетной сессии для систематизации знаний.

7. Оценочные средства

7.1. Паспорт оценочных средств

Семестр	Код контролируемой компетенции (или ее части)	Наименование оценочного средства
7	ПК-3	Тестовые задания 1 - 300 Вопросы к зачету 1 – 70 Отчеты по практическим работам 1-8

7.2. Типовые задания или иные материалы, необходимые для текущего контроля

7.2.1. Типовые тестовые материалы

(наименование оценочного средства)

1. Какие из перечисленных моделей считаются изначально интерпретируемыми?

- A. Линейная регрессия
- B. Глубокая нейронная сеть с 50 слоями
- C. Дерево решений глубиной 3
- D. Ансамбль из 1000 случайных деревьев

Ответ: A, C

2. Какие утверждения о компромиссе между точностью и интерпретируемостью верны?

- A. Сложные модели всегда точнее и интерпретируемее простых.
- B. Простые модели часто более интерпретируемы, но могут проигрывать в точности.
- C. Интерпретируемость всегда достигается исключительно за счет снижения точности.

D. Выбор модели часто представляет собой баланс между точностью и интерпретируемостью.

Ответ: B, D

3. Что из перечисленного является целью методов интерпретируемого ИИ (XAI)?

- A. Обеспечение соответствия нормативным требованиям (например, GDPR)
- B. Ускорение времени обучения моделей
- C. Отладка и улучшение моделей
- D. Формирование доверия у пользователей

Ответ: A, C, D

4. Какие методы относятся к модельно-агностическим (model-agnostic)?

- A. Анализ весов в линейной регрессии

- B. Алгоритм LIME
 - C. Алгоритм SHAP
 - D. Визуализация дерева решений
- Ответ: B, C

5. Какие утверждения о локальной интерпретации верны?
- A. Она объясняет общее поведение модели на всем наборе данных.
 - B. LIME является типичным представителем методов локальной интерпретации.
 - C. Она отвечает на вопрос "Почему модель приняла такое решение для этого конкретного примера?"
 - D. Она всегда требует полного переобучения модели.
- Ответ: B, C

6. Какие из перечисленных методов основаны на идее суррогатной модели?
- A. Permutation Importance
 - B. LIME
 - C. Global Surrogate
 - D. SHAP (KernelExplainer)
- Ответ: B, C, D

7. Какие визуализации являются стандартными для анализа глобальной важности признаков?
- A. График важности признаков (Feature Importance Bar Chart)
 - B. Summary plot в SHAP
 - C. Force plot для одного наблюдения
 - D. График частичных зависимостей (PDP)
- Ответ: A, B

8. Какие методы используются для интерпретации решений сверточных нейронных сетей (CNN) в компьютерном зрении?
- A. Дерево решений
 - B. Grad-CAM
 - C. Карты активации (Activation Maps)
 - D. Линейная регрессия
- Ответ: B, C

9. Какие утверждения о методе SHAP верны?
- A. Он основан на теории игр и концепции справедливого распределения вклада.
 - B. Он может быть рассчитан только для линейных моделей.
 - C. Он предоставляет как глобальные, так и локальные объяснения.
 - D. Его вычисление всегда требует экспоненциального времени.
- Ответ: A, C

10. Какие проблемы призван решать метод LIME?
- A. Низкая точность сложных моделей
 - B. Интерпретация индивидуальных предсказаний моделей-"черных ящиков"
 - C. Высокая вычислительная сложность методов на основе Шепли
 - D. Отсутствие встроенной интерпретируемости у многих моделей
- Ответ: B, D

Критерии оценки за пройденный тест:

- 40 баллов выставляется обучающемуся, если он ответил правильно на все вопросы

- рандомной выборки 30 тестовых заданий;
- 0-39 баллов выставляется обучающемуся в зависимости от количества верных ответов на вопросы рандомной выборки 30 тестовых заданий.

7.2.2. Пример практической работы

Практическая работа 1. Анализ встроенной интерпретируемости линейных моделей и деревьев

Цель работы: освоить извлечение и анализ встроенных метрик интерпретируемости (веса коэффициентов, важность признаков) из простых моделей.

Задание. На имеющемся наборе данных построить линейную регрессию и дерево решений. Проанализировать и визуализировать веса коэффициентов и важность признаков, сделав выводы о влиянии факторов на целевую переменную.

Методические указания:

1. Загрузите данные и выполните первичный анализ (head, describe, info).
2. Разбейте данные на обучающую и тестовую выборки.
3. Обучите модель линейной регрессии.
4. Визуализируйте коэффициенты модели с помощью столбчатой диаграммы.
5. Обучите модель дерева решений с ограниченной глубиной.
6. Визуализируйте важность признаков и постройте схему самого дерева.
7. Сравните выводы, полученные от двух моделей.
8. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 2. Расчет и визуализация глобальной важности признаков

Цель работы: научиться вычислять и сравнивать глобальную важность признаков с помощью модельно-агностических методов (на примере Permutation Importance).

Задание. Обучить модель случайного леса на предоставленных данных. Рассчитать и отранжировать важность признаков методом перестановок, проанализировав стабильность результатов.

Методические указания:

1. Обучите модель случайного леса.
2. Рассчитайте встроенную важность признаков (feature_importances_).
3. Используя библиотеку ELI5 или scikit-learn, рассчитайте важность признаков методом перестановок (Permutation Importance).
4. Визуализируйте и сравните результаты двух методов.
5. Проанализируйте, какие признаки модель считает наиболее влиятельными.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 3. Применение LIME для объяснения отдельных прогнозов

Цель работы: приобрести навыки локальной интерпретации произвольной модели-«черного ящика» с использованием библиотеки LIME.

Задание. Для обученной модели (например, градиентного бустинга) и нескольких тестовых примеров (корректный и ошибочный прогноз) сгенерировать и проанализировать локальные объяснения с помощью LIME.

Методические указания:

1. Выберите два примера из тестовой выборки: где модель предсказала верно и где ошиблась.
2. Инициализируйте объяснитель LIME для табличных данных.
3. Сгенерируйте объяснение для каждого выбранного примера.
4. Визуализируйте результат (show_in_notebook).

5. Проанализируйте, какие признаки наиболее сильно повлияли на конкретное предсказание в каждом случае.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 4. Применение SHAP для анализа моделей и прогнозов

Цель работы: освоить использование фреймворка SHAP для проведения как локального, так и глобального анализа моделей.

Задание. Провести всесторонний анализ модели градиентного бустинга с помощью SHAP: рассчитать глобальную важность признаков, проанализировать индивидуальные предсказания и зависимости.

Методические указания:

1. Рассчитайте SHAP-значения для тестовой выборки.
2. Постройте график сводной важности признаков (summary_plot).
3. Постройте график зависимости (dependence_plot) для одного-двух наиболее важных признаков.
4. Используя force_plot, проанализируйте вклад признаков в прогноз для конкретного примера.
5. Сравните выводы SHAP с результатами, полученными в работах 2 и 3.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 5. Визуализация карт активации и атрибуции в нейронных сетях

Цель работы: научиться применять методы атрибуции (Grad-CAM) для визуализации областей интереса в изображениях у сверточной нейронной сети (CNN).

Задание. Для предобученной CNN (например, VGG) реализовать метод Grad-CAM для визуализации участков изображения, наиболее повлиявших на итоговую классификацию.

Методические указания:

1. Загрузите предобученную модель и пример изображения.
2. Выберите целевой сверточный слой для анализа.
3. Реализуйте алгоритм Grad-CAM: рассчитайте градиенты целевого класса по активациям выбранного слоя.
4. Наложите полученную карту атрибуции на исходное изображение.
5. Сделайте выводы о том, на какие части изображения модель обратила внимание.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 6. Анализ внимания в моделях для обработки текста

Цель работы: исследовать работу механизма внимания в нейронной сети для анализа тональности текста и интерпретации предсказаний.

Задание. Построить или использовать готовую модель с механизмом внимания для классификации тональности отзывов. Проанализировать веса внимания для ключевых слов в различных примерах.

Методические указания:

1. Загрузите набор данных с текстовыми отзывами.
2. Обучите или загрузите простую модель с слоем Attention.
3. Получите веса внимания для нескольких тестовых примеров (положительных и отрицательных отзывов).
4. Визуализируйте результат, окрашивая слова в тексте в соответствии с их весом внимания.
5. Объясните, как модель пришла к решению на основе ключевых слов.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 7. Сравнительный анализ методов ХАІ на комплексном кейсе

Цель работы: закрепить навыки выбора и применения различных методов ХАІ для решения сквозной задачи и формирования итогового заключения.

Задание. На реальном кейсе (например, "Кредитный скоринг") провести полный анализ модели с использованием 3-4 различных методов ХАІ (SHAP, LIME, Permutation Importance), подготовив сводный отчет.

Методические указания:

1. Обучите модель для решения задачи бинарной классификации.
2. Последовательно примените несколько изученных методов интерпретации.
3. Сравните результаты, полученные разными методами: сходства и противоречия.
4. Сформулируйте итоговое заключение о том, как модель принимает решения, и насколько она надежна.
5. Определите, какой метод лучше подходит для объяснения бизнес-аналитику, а какой — для разработчика.
6. Составьте отчет, содержащий все пункты выполнения задания.

Практическая работа 8. Разработка тестового сценария для верификации объяснений модели

Цель работы: сформировать навыки тестирования программных компонентов, связанных с интерпретируемостью, в соответствии с компетенцией ПК-3.

Задание. Разработать тестовый сценарий (test case) для проверки корректности работы метода интерпретации (на выбор: LIME или SHAP) на синтетических и реальных данных.

Методические указания:

1. Определите критерии качества объяснения: устойчивость (robustness) и непротиворечивость (faithfulness).
2. Подготовьте синтетический датасет с заведомо известными зависимостями.
3. Напишите тест, проверяющий, что метод ХАІ корректно идентифицирует эти зависимости.
4. Напишите тест, проверяющий, что небольшое изменение входных данных не приводит к кардинальному изменению объяснения (для устойчивых методов).
5. Оформите тестовые сценарии в виде кода (например, с использованием pytest).
6. Составьте отчет, содержащий все пункты выполнения задания.

Требования к оформлению

Отчет должен содержать подробное описание (включая иллюстрации). Отчёт по практическому занятию выполняется на страницах формата А4 в электронном виде.

При оформлении отчёта используется сквозная нумерация страниц, считая титульный лист первой страницей. Номер страницы на титульном листе не ставится. Номера страницы ставятся по центру сверху.

При оформлении отчёта соблюдать следующие требования:

- Для заголовков: полужирный шрифт, 14 пт, центрированный.
- Для основного текста: нежирный шрифт, 14 пт, выравнивание по ширине.
- Во всех случаях тип шрифта – Times New Roman, отступ абзаца 1.25 см, полуторный междустрочный интервал.
- Поля: левое – 2 см, правое, верхнее и нижнее – 1 см.

Процедура оценивания

Оценка выполненной практической работы проводится по следующим критериям:

1. Наличие всей существенной информации по работе

2. Точность и полнота предоставляемых сведений
3. Непротиворечивость приводимой информации
4. Правильность интерпретаций и выводов, которые сделаны по результатам работы
5. Степень достижения обучающимся поставленной цели
6. Обоснованность применяемого решения
7. Грамотность (содержательная) используемых формулировок

Критерии оценки:

- оценка «зачтено» выставляется обучающемуся, если
 - продемонстрирована работа программы;
 - предоставлен отчет о выполнении работы, оформленный в соответствии с установленными требованиями;
 - при защите отчета продемонстрированы всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений, понимание и умение объяснить код программы;
- оценка «не зачтено» выставляется обучающемуся, если
 - продемонстрирована работа программы, не соответствующей заданию;
 - не предоставлен отчет о выполнении работы, оформленный в соответствии с установленными требованиями;
 - при защите отчета не продемонстрированы знания учебной программы дисциплины, не наблюдается понимание кода программы;

Критерии оценки за отчеты по практическим работам:

Формы текущего контроля	Критерии и нормы оценки
Отчет по практическим работам 1-6	7 баллов – задание выполнено в полном объеме без замечаний 4-6 баллов – задание выполнено не в полном объеме, присутствуют несущественные замечания 1-3 балла – задание выполнено не в полном объеме, присутствуют замечания по выполнению задания 0 баллов – задание не выполнено
Отчет по практическим работам 7-8	9 баллов – задание выполнено в полном объеме без замечаний 6-7 баллов – задание выполнено в полном объеме, присутствуют замечания по выполнению задания 4-5 баллов – задание выполнено не в полном объеме, присутствуют несущественные замечания 1-3 балла – задание выполнено не в полном объеме, присутствуют замечания по выполнению задания 0 баллов – задание не выполнено

7.3. Оценочные средства для промежуточной аттестации по итогам освоения дисциплины

7.3.1. Вопросы к промежуточной аттестации

Семестр _____ 7 _____

№	Вопросы к зачету
1.	Дайте определение интерпретируемого искусственного интеллекта (ХАИ).
2.	В чем заключается фундаментальная проблема «черного ящика» в машинном обучении?
3.	Назовите и охарактеризуйте два основных типа интерпретируемости: внутренняя (inherent) и пост-хок (post-hoc).
4.	В чем разница между глобальной и локальной интерпретацией?
5.	Перечислите основные причины, по которым интерпретируемость моделей ИИ стала критически важной.
6.	Опишите компромисс между точностью и интерпретируемостью модели.
7.	Назовите категории стейкхолдеров, заинтересованных в интерпретируемости, и их цели.
8.	Каковы этические аспекты использования интерпретируемого ИИ?
9.	Дайте определение «справедливости» (fairness) в контексте машинного обучения и роль ХАИ в ее обеспечении.
10.	Что такое «устойчивость» (robustness) модели и как ее можно проверить с помощью методов ХАИ?
11.	Какие модели машинного обучения считаются изначально интерпретируемыми и почему?
12.	Опишите, как интерпретировать коэффициенты в модели линейной регрессии.
13.	Как можно визуализировать и интерпретировать модель дерева решений?
14.	В чем преимущества и недостатки логистической регрессии как интерпретируемой модели?
15.	Что такое обобщенные аддитивные модели (GAM) и в чем их интерпретируемость?
16.	Почему правило IF-THEN считается интерпретируемым?
17.	Какие ограничения накладываются на изначально интерпретируемые модели для сохранения их понятности?
18.	Что подразумевается под «важностью признака» (feature importance)?
19.	Опишите принцип работы метода перестановочной важности (Permutation Importance).
20.	В чем разница между встроенной важностью в Random Forest и перестановочной важностью?
21.	Что такое частичные зависимости (Partial Dependence Plots - PDP)?
22.	Как интерпретируется график частичных зависимостей?
23.	Что показывают графики индивидуальных условных ожиданий (ICE)?
24.	В чем преимущества и недостатки ALE (Accumulated Local Effects) plots по сравнению с PDP?
25.	Что такое анализ профилей предсказаний (Prediction Profile)?
26.	Опишите основную идею алгоритма LIME.
27.	Каковы ключевые этапы работы LIME для табличных данных?
28.	Как LIME генерирует новые примеры в окрестности объясняемого объекта?
29.	В чем слабые стороны подхода LIME?
30.	Как выбирается суррогатная модель в LIME и почему?
31.	Опишите, как LIME можно применить для интерпретации текстовых данных.
32.	Опишите, как LIME можно применить для интерпретации изображений.
33.	Объясните, что такое значения Шепли (Shapley Values) из теории игр.
34.	Как значения Шепли применяются для интерпретации моделей машинного

№	Вопросы к зачету
	обучения?
35.	Дайте определение силы Шепли для отдельного признака.
36.	В чем заключаются вычислительные сложности точного расчета значений Шепли?
37.	Какие существуют аппроксимации для вычисления SHAP-значений (KernelSHAP, TreeSHAP)?
38.	Объясните, что показывает summary plot в SHAP.
39.	Как интерпретируется force plot для индивидуального предсказания?
40.	Что визуализирует dependence plot в SHAP?
41.	В чем преимущества SHAP перед LIME?
42.	Дайте определение суррогатной модели в контексте XAI.
43.	Какой основной принцип использования суррогатных моделей для интерпретации?
44.	Назовите типичные алгоритмы, используемые в качестве суррогатных моделей.
45.	Опишите процесс построения и использования глобальной суррогатной модели.
46.	Каковы критерии качества суррогатной модели?
47.	В чем ограничения подхода, основанного на суррогатных моделях?
48.	Какие специфические challenges возникают при интерпретации глубоких нейронных сетей?
49.	Что такое карты активации (Activation Maps) и как они используются для интерпретации?
50.	Опишите принцип работы метода Grad-CAM.
51.	В чем разница между CAM и Grad-CAM?
52.	Что такое Guided Backpropagation и Deconvolution?
53.	Что такое Integrated Gradients?
54.	Как методы атрибуции объясняют вклад отдельных пикселей в предсказание?
55.	Что такое "Attention Mechanisms" и как они обеспечивают интерпретируемость?
56.	Как визуализируются карты внимания в моделях для обработки изображений?
57.	Как интерпретируются веса внимания в моделях для обработки текста?
58.	Каковы основные подходы к интерпретации моделей в компьютерном зрении?
59.	Как визуализируются наиболее важные области изображения для предсказания модели?
60.	Какие методы используются для объяснения решений моделей, работающих с текстом?
61.	Что такое LIME для текста и как он работает?
62.	Как SHAP применяется для анализа текстовых классификаторов?
63.	Как можно интерпретировать модели машинного перевода или текстовых генераторов?
64.	В чем особенности интерпретации рекомендательных систем?
65.	Какие существуют метрики для оценки качества интерпретаций?
66.	Что такое «непротиворечивость» (faithfulness) объяснений?
67.	Что такое «устойчивость» (robustness) объяснений?
68.	Как можно проверить надежность метода интерпретации?
69.	Какие существуют лучшие практики для визуализации результатов XAI?
70.	Как следует представлять результаты интерпретации нетехническим стейкхолдерам?

7.3.2. Критерии и нормы оценки

Семестр	Форма проведения промежуточной аттестации	Критерии и нормы оценки	
7	Зачет (по накопительному рейтингу)	«зачтено»	рейтинговый балл 55-100
		«не зачтено»	рейтинговый 0-54

8. Учебно-методическое и информационное обеспечение дисциплины

8.1. Обязательная литература

№ п/п	Авторы, составители	Заглавие (заголовок)	Тип (учебник, учебное пособие, учебно-методическое пособие, практикум, др.)	Год издания	Количество в научной библиотеке / Наименование ЭБС
1	Л. С. Болотова	Системы искусственного интеллекта: модели и технологии, основанные на знаниях : учебник / Л. С. Болотова. - Москва : Финансы и статистика, 2023. - 666 с. - ISBN 978-5-00184-097-8.	Учебник	2023	ЭБС «ZNANIUM»
2	А.В. Андрейчиков, О.Н. Андрейчикова	Интеллектуальные информационные системы и методы искусственного интеллекта : учебник / А.В. Андрейчиков, О.Н. Андрейчикова. — Москва : ИНФРА-М, 2025. — 530 с. + Доп. материалы [Электронный ресурс]. — (Высшее образование). — DOI 10.12737/1009595. - ISBN 978-5-16-020880-0.	Учебник	2025	ЭБС «ZNANIUM»
3	С.О. Крамаров	Искусственный интеллект в образовании: возможности, методы и рекомендации для педагогов : учебно-практическое пособие / под ред. С.О. Крамарова. — Москва : РИОР : ИНФРА-М, 2026. — 99 с. — (Наука и практика). — DOI: https://doi.org/10.29039/02147-7 . - ISBN 978-5-369-01968-9	Учебно-практическое пособие	2026	ЭБС «ZNANIUM»

8.2. Дополнительная литература

№ п/п	Авторы, составители	Заглавие (заголовок)	Тип (учебник, учебное пособие, учебно- методическое пособие, практикум, др.)	Год издания	Количество в научной библиотеке / Наименование ЭБС
1	А. И. Пиляй, Л. А. Адамцевич	Основы методов искусственного интеллекта : учебно-методическое пособие / А. И. Пиляй, Л. А. Адамцевич. — Москва : МИСИ-МГСУ, ЭБС АСВ, 2023. — 60 с. — ISBN 978-5-7264-3307-3.	Учебное пособие	2023	ЭБС «IPRBooks»
2	С. Л. Сотник	Проектирование систем искусственного интеллекта : учебное пособие / С. Л. Сотник. — 4-е изд. — Москва : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2025. — 228 с. — ISBN 978-5-4497-0868-7.	Учебное пособие	2025	ЭБС «IPRBooks»
3	А. А. Алетдинова, М. Г. Гриф	Системы искусственного интеллекта : учебное пособие / А. А. Алетдинова, М. Г. Гриф. — Новосибирск : Новосибирский государственный технический университет, 2023. — 76 с. — ISBN 978-5-7782-5124-3.	Учебное пособие	2023	ЭБС «IPRBooks»

8.3. Перечень профессиональных баз данных и информационных справочных систем

№ пп	Наименование	Ссылка
1	Springer Nature (Полнотекстовая коллекция журналов)	https://www.springernature.com/gp/products
2	Springer eBooks (Полнотекстовая коллекция электронных книг издательства Springer Nature)	https://link.springer.com/
3	«Кодекс»	https://kodeks.ru/
4	ELIBRARY.RU (электронная библиотека научных публикаций)	http://elibrary.ru
5	"Гарант"	https://www.garant.ru/
6	"КонсультантПлюс"	https://www.consultant.ru/
7	Техэксперт	https://cntd.ru/

8.4. Перечень программного обеспечения

№ п/п	Наименование ПО	Реквизиты лицензии / договора
1	Python 3.x	Лицензия: Python Software Foundation License (свободное ПО)
2	PyTest (framework для автоматизации тестирования)	Лицензия: MIT License (свободное ПО)
3	Postman (инструмент для тестирования API)	Бесплатная версия для образовательного и личного использования
4	Jupyter Notebook / JupyterLab	Лицензия: BSD License (свободное ПО)
5	Scikit-learn (библиотека ML для расчёта метрик)	Лицензия: BSD License
6	Pandas / NumPy	Лицензия: BSD License
7	Git	Лицензия: GNU General Public License (GPL, свободное ПО)
8	GitHub (web-интерфейс)	Бесплатный доступ для образовательного использования (Free tier)
9	VS Code (редактор кода)	Лицензия: MIT License (бесплатно)
10	Selenium (по необходимости)	Лицензия: Apache License 2.0

8.5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

№ п/п	Наименование оборудованных учебных кабинетов, лабораторий, мастерских и др. объектов для проведения практических и лабораторных занятий, помещений для самостоятельной работы обучающихся (номер аудитории)	Перечень основного оборудования
1	Компьютерный класс. Учебная аудитория для проведения занятий лекционного типа.	Компьютер (монитор 19", системный блок Pentium (R) Dual-Core E5500 2,8

№ п/п	Наименование оборудованных учебных кабинетов, лабораторий, мастерских и др. объектов для проведения практических и лабораторных занятий, помещений для самостоятельной работы обучающихся (номер аудитории)	Перечень основного оборудования
	Учебная аудитория для проведения занятий семинарского типа. Учебная аудитория для проведения лабораторных работ. Учебная аудитория для курсового проектирования (выполнения курсовых работ). Учебная аудитория для проведения групповых и индивидуальных консультаций. Учебная аудитория для проведения занятий текущего контроля и промежуточной аттестации. (УЛК-401)	GHz / 4 Gb / 500 Gb), столы ученические, столы компьютерные, стол преподавательский, стулья, доска аудиторная (меловая).
2	Помещение для самостоятельной работы обучающихся (УЛК-105)	Стол, стулья, стеллажи (в т.ч. выставочные) с книгами, компьютеры, мобильные рабочие места.
3	Помещение для самостоятельной работы обучающихся (УЛК-406)	Стол компьютерный, стулья, микрокомпьютеры raspberry pi 32 bit