МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ – МСХА имени К.А. ТИМИРЯЗЕВА»

(ФГБОУ ВО РГАУ - МСХА имени К.А. Тимирязева)

Институт экономики и управления АПК
Кафедра статистики и кибернетики

УТВЕРЖДАЮ:
Директор института
экономики и управления АПК
_____ Л.И. Хоружий

«28» августа 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Б1.В.02 Наука о данных (Data Science)

для подготовки магистров

ФГОС ВО

Направление: 09.04.02 Информационные системы и технологии
Направленность: Науки о данных

Курс 1
Семестр 2

Форма обучения: очная

Год начала подготовки: 2025

Москва, 2025

Разработчики: Калитвин В.А., канд. ф.-м. наук, доцент
(ФИО, ученая степень, ученое звание) (подпись)

«26» августа 2025 г.

Рецензент: Прудкий А.С., к.пед.н., доцент
(ФИО, ученая степень, ученое звание) (подпись)

«26» августа 2025 г.

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.04.02 Информационные системы и технологии, профессионального стандарта и учебного плана 2025 года начала подготовки.

Программа обсуждена на заседании кафедры статистики и кибернетики протокол № 11 от «26» августа 2025 г.

И.о. зав. кафедрой Уколова А.В., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание) (подпись)

«26» августа 2025 г.

Согласовано:

Председатель учебно-методической
комиссии института экономики и управления АПК
Гупалова Т.Н., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание) (подпись)

«26» августа 2025 г.

И.о. зав. выпускающей кафедрой статистики и кибернетики
Уколова А.В., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание) (подпись)

«28» августа 2025 г.

Заведующий отделом комплектования ЦНБ

СОДЕРЖАНИЕ

АННОТАЦИЯ

рабочей программы учебной дисциплины Б1.В.02 «Наука о данных (Data Science)» для подготовки магистров по направлению 09.04.02 Информационные системы и технологии, направленности «Науки о данных»

Цель освоения дисциплины:

сформировать у обучающихся целостное понимание теоретических основ и прикладных аспектов науки о данных, а также развить практические умения и нывыки решения реальных задач анализа данных.

По окончанию изучения дисциплины студент должен знать: базовые понятия и терминологию науки о данных (Data Science, DS) - обучение с учителем и без учителя, классификацию задач машинного обучения, понятие признака, объекта, целевой переменной, этапы жизненного цикла ML-проекта (от постановки задачи до создания готового программного продукта); основные классы алгоритмов и их принципы работы - линейная и логистическая регрессия, регуляризация, деревья решений и ансамблевые методы, метрические методы, методы кластеризации, метод опорных векторов и ядра, базовые архитектуры нейронных сетей (полносвязные, свёрточные), методы снижения размерности; методологию подготовки и анализа данных (способы обработки пропусков, выбросов, категориальных признаков, техники масштабирования и нормализации, принципы отбора и конструирования признаков, методы работы с несбалансированными данными); подходы к оценке качества моделей, принципы кросс-валидации и разбиения данных (train/val/test); способы настройки и оптимизации моделей (подбор гиперпараметров, борьба с переобучением); инструменты разработки и развертывания моделей; современные тренды и ограничения машинного обучения.

Место дисциплины в учебном плане: дисциплина включена в часть, формируемую участниками образовательных отношений учебного плана по направлению подготовки 09.04.02 Информационные системы и технологии.

Требования к результатам освоения дисциплины: в результате освоения дисциплины формируются следующие компетенции (индикаторы): ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3) .

Краткое содержание дисциплины:

Основы машинного обучения. Подготовка данных. Визуализация данных. Отбор признаков. Обучение с учителем. Регрессия. Классификация. Ансамбли моделей. Поиск ассоциативных правил в процессе анализа данных. Обучение без учителя. Кластерный анализ. Анализ текстовой информации и аналитика.

Общая трудоемкость дисциплины: 72/ 2 (часы/зач. ед.).
Промежуточный контроль: зачет с оценкой.

1. Цель освоения дисциплины

Целью дисциплины «Наука о данных (Data Science)» является формирование у обучающихся целостное понимание теоретических основ и прикладных аспектов науки о данных, а также развитие практических умений и нывыков решения реальных задач анализа данных. По окончанию изучения дисциплины студент должен знать: базовые понятия и терминологию науки о данных (Data Science, DS) - обучение с учителем и без учителя, классификацию задач машинного обученияй, понятие признака, объекта, целевой переменной, этапы жизненного цикла ML-проекта (от постановки задачи до создания готового про-

граммного продукта); основные классы алгоритмов и их принципы работы - линейная и логистическая регрессия, регуляризация, деревья решений и ансамблевые методы, метрические методы, методы кластеризации, метод опорных векторов и ядра, базовые архитектуры нейронных сетей (полносвязные, свёрточные), методы снижения размерности; методологию подготовки и анализа данных (способы обработки пропусков, выбросов, категориальных признаков, техники масштабирования и нормализации, принципы отбора и конструирования признаков, методы работы с несбалансированными данными); подходы к оценке качества моделей, принципы кросс-валидации и разбиения данных (train/val/test); способы настройки и оптимизации моделей (подбор гиперпараметров, борьба с переобучением); инструменты разработки и развертывания моделей; современные тренды и ограничения машинного обучения.

2. Место дисциплины в учебном процессе

Дисциплина «Наука о данных (Data Science)» включена в часть учебного плана, формируемую участниками образовательных отношений. Дисциплина «Наука о данных (Data Science)» реализуется в соответствии с требованиями ФГОС ВО, ОПОП ВО и Учебного плана по направлению 09.04.02 Информационные системы и технологии.

Дисциплина «Наука о данных (Data Science)» изучается на первом курсе образовательного цикла.

Для успешного изучения дисциплины необходимы знания и умения по предшествующим дисциплинам:

«Специальные главы математики», «Модели информационных процессов и систем», «Статистика (продвинутый уровень)», «Эконометрика (продвинутый уровень)», «Инструменты Data Science в R, Python, SQL».

Дисциплина «Наука о данных (Data Science)» является основополагающей для изучения следующих дисциплин: «Системы поддержки принятия решений», «Системы искусственного интеллекта», «Глубокое обучение в науках о данных в сельском хозяйстве», «Анализ больших данных в сельском хозяйстве», «Компьютерное зрение в сельском хозяйстве».

Особенностью дисциплины является реализация алгоритмов машинного обучения средствами языка программирования Python.

Рабочая программа дисциплины «Наука о данных (Data Science)» для инвалидов и лиц с ограниченными возможностями здоровья разрабатывается индивидуально с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся.

3. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Изучение дисциплины направлено на формирование у обучающихся профессиональных компетенций, представленных в таблице 1.

Таблица 1

Требования к результатам освоения учебной дисциплины «Наука о данных (Data Science)»

№ п/п	Код компетенции	Содержание компетенции (или её части)	Индикаторы компетенций	В результате изучения учебной дисциплины обучающиеся должны:		
				знать	уметь	владеть
1	ПКос-2	Способность проводить анализ данных с использованием информационных технологий в области сельского хозяйства, экономики, бухгалтерского учета, статистики, финансов и др.	ПКос-2.1 Знать: основы технологии производства продукции сельского хозяйства; теорию и методологию дисциплин экономического профиля (экономика, бухгалтерский учет, статистика, финансы и др.); информационные технологии анализа данных; источники информации для профессиональной деятельности	информационные технологии анализа данных с применением машинного обучения; источники информации для профессиональной деятельности		
			ПКос-2.2 Уметь: собирать информацию для проведения анализа данных в области сельского хозяйства, экономики, бухгалтерского учета, статистики, финансов и др.; устанавливать причинно-следственные связи между признаками; выбирать и		собирать информацию для проведения анализа данных в области сельского хозяйства, экономики, бухгалтерского учета, статистики, финансов и др.; устанавливать причинно-следственные связи между признаками; выбирать и применять, в том числе с использованием современных информационных технологий и алгоритмов машинного обучения, методы анализа данных	

			применять, в том числе с использованием современных информационных технологий, методы анализа данных в области сельского хозяйства, экономики, бухгалтерского учета, статистики, финансов и др.; делать выводы на основе проведенного анализа данных		в области сельского хозяйства, экономики, бухгалтерского учета, статистики, финансов и др.; делать выводы на основе проведенного анализа данных	
			ПКос-2.3 Владеть: методологией и навыками проведения анализа данных с использованием информационных технологий в области сельского хозяйства, в том числе экономики сельского хозяйства			методологией и навыками проведения анализа данных с использованием информационных технологий и методов машинного обучения в области сельского хозяйства
2	ПКос-3	Способен совершенствовать и разрабатывать новые методы, модели, алгоритмы, технологии и инструментальные средства работы с данными, в т.ч. большими данными в сельском хозяйстве	ПКос-3.1 Знать: методы науки о данных, в т.ч. методы машинного обучения, обработки и визуализации больших данных; состояние и перспективы развития - науки о данных, ис-	методы науки о данных, в т.ч. методы машинного обучения, обработки и визуализации больших данных; состояние и перспективы развития - науки о данных, используемого при обработке данных программного инструментария; потребности		

			пользуемого при обработке данных программного инструментария; потребности в совершенствовании и разработке новых методов, технологий и инструментальных средств для работы с данными, в т.ч. большими; область применения науки о данных в сельском хозяйстве	в совершенствовании и разработке новых методов, технологий и инструментальных средств для работы с данными, в т.ч. большими; область применения науки о данных в сельском хозяйстве		
			ПКос-3.2 Уметь: определять перспективную тематику научно-исследовательских работ в области совершенствования и разработки новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с данными; планировать и проводить аналитические и научные исследования по тематике информационных технологий в АПК, применяемых		определять перспективную тематику научно-исследовательских работ в области совершенствования и разработки новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с данными; планировать и проводить аналитические и научные исследования по тематике информационных технологий в АПК, применяемых в науке о данных	

			в науке о данных			
			ПКос-3.3 Иметь навыки: разработки новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с данными на основе анализа потребностей и передового зарубежного и отечественного опыта; планирования состава и содержания, согласование перечня научно-исследовательских работ в профессиональной деятельности в АПК			разработки новых методов, моделей, алгоритмов, технологий и инструментальных средств работы с данными на основе анализа потребностей и передового зарубежного и отечественного опыта; планирования состава и содержания, согласование перечня научно-исследовательских работ в профессиональной деятельности в АПК

4. Структура и содержание дисциплины

4.1 Распределение трудоёмкости дисциплины по видам работ по семестрам

Общая трудоёмкость дисциплины составляет 2 зач.ед. (72 часа), их распределение по видам работ и семестрам представлено в таблице 2.

Таблица 2

Распределение трудоёмкости дисциплины по видам работ во 2 семестре

Вид учебной работы	Трудоёмкость	
	час. всего/*	в т.ч. по семестрам
		№ 2
Общая трудоёмкость дисциплины по учебному плану	**72**	**72**
1. Контактная работа:	**30,35**	**30,35**
Аудиторная работа	**30**	**~~30~~**
лекции (Л)	**-**	-
практические занятия (ПЗ)	**30/4**	30/4
контактная работа на промежуточном контроле (КРА)	**0,35**	0,35
2. Самостоятельная работа (СРС)	**41,65**	**41,65**
самостоятельное изучение разделов, самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к практическим занятиям)	41,65	41,65
Вид промежуточного контроля:	Зачет с оценкой	

* в том числе практическая подготовка

4.2 Содержание дисциплины

Таблица 3

Тематический план учебной дисциплины

Наименование разделов и тем дисциплин (укрупнённо)	Всего	Аудиторная работа			Внеаудиторная работа СР
		Л	ПЗ всего/*	ПКР	
Тема 1. Основы науки о данных (DS)	8,65	-	2	-	6,65
Тема 2. Предобработка данных и их визуализация	9	-	4	-	5
Тема 3. Отбор признаков	9	-	4	-	5
Тема 4. Обучение с учителем	9	-	4/2	-	5
Тема 5. Поиск ассоциативных правил в процессе анализа данных	9	-	4	-	5
Тема 6. Кластерный анализ	9	-	4/2	-	5
Тема 7. Нейронные сети	9	-	4		5
Тема 8. Анализ текстовой информации и аналитика	9	-	4	-	5

Наименование разделов и тем дисциплин (укрупнённо)	Всего	Аудиторная работа			Внеаудиторная работа СР
		Л	ПЗ всего/*	ПКР	
Контактная работа на промежуточном контроле (КРА)	0,35	-	-	0,35	-
Всего за 2 семестр	**72**	**-**	**30**	**0,35**	**41,65**
Итого по дисциплине	**72**	**-**	**30**	**0,35**	**41,65**

Тема 1. Основы науки о данных (Data Science).

История развития науки о данных. Основные понятия. Примеры задач и областей приложения. Типы задач машинного обучения. Классификация методов машинного обучения. Классификация. Регрессия. Типы ошибок классификации. Обобщающая способность классификатора. Формализация и постановка задачи машинного обучения. Недообучение. Переобучение.

Тема 2. Предобработка данных и их визуализация.

Обнаружение пропущенных данных. Классификация типов пропущенных данных. Исследование структуры пропущенных данных Визуализация закономерностей в пропущенных данных. Анализ полных наблюдений. Множественное восстановление пропущенных данных. Методы и средства визуального представления информации, в частности, способы представления информации в одно-, двух-, трехмерном измерениях, а также способы отображения информации в более чем трех измерениях.

Описание принципов качественной визуализации. Основные тенденции в области визуализации.

Тема 3. Отбор признаков.

Поиск дубликатов и противоречии во входных данных. Корреляционный анализ. Коэффициент корреляции Пирсона. Поиск взаимосвязей. Отбор факторов и снижение размерности исходных данных. Факторный анализ. Метод главных компонент. Структура и шум в данных..

Тема 4. Обучение с учителем.

Правила классификации и методы их построения. Построение основных методов классификации в Python. Математические основы работы алгоритмов. Метрики качества задач классификации. Дерево решений. Метод опорных векторов. Метод опорных векторов с ядерной функцией. Случайный лес. Логистическая регрессия. Дискриминантный анализ. Байесовская (наивная) классификация. Метод ближайшего соседа. ROC-кривые. Проверка классификатора. Проверка тестовой выборкой. Перекрестная проверка. Оценка информативности признаков. Евклидово расстояние. Расстояние Махалонобиса.
Чувствительность и избирательность. Кривая мощности критерия классификации. Бустинг и переобучение. Параллельные методы комитетов:
бутстреп, бэггинг. Корреляционные и причинно-следственные связи. Корреляция признаков и структура данных. Латентные структуры в данных. Метод

наименьших квадратов. Теорема ГауссаМаркова. Обобщенный метод наименьших квадратов. Рекурсивный метод наименьших квадратов. Анализ регрессионных остатков. Формальная и эффективная размерность. Графическая проверка линейности, гомоскедастичности. Объясненная и необъясненная вариация. Коэффициент детерминации. Неустойчивость МНК к выбросам. Робастная регрессия. Метрики качества задач регрессии. Множественная линейная регрессия, ее преимущества и недостатки. Мультиколлинеарность данных. Метод главных компонент.

Тема 5. Поиск ассоциативных правил в процессе анализа данных.

Понятия и методы выявления закономерностей в интеллектуальном анализе данных. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование). Анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.

Тема 6. Кластерный анализ.

Иерархические и неиерархические методы в кластерном анализе. Рассмотрение примеров использования кластерного анализа. Кластеризация как классификация без учителя. Меры сходства и меры различия образов. Критерии качества кластеризации. Итеративная оптимизация разбиения на кластеры. Плоские методы кластеризации. Метод K средних. Метод ISODATA. Метод FOREL. Графовые методы. Иерархическая кластеризация. Агломеративные и разделяющие алгоритмы кластеризации. Дендрограммы.

Тема 7. Нейронные сети.

Введение в нейронные сети. Краткая история развития (от перцептрона 1958 г. до глубокого обучения). Устройство искусственного нейрона. Архитектура нейронной сети. Многослойные нейронные сети. Задачи, решаемые нейронными сетями. Визуализация работы нейрона. Обучение нейронных сетей. Метод обратного распространения ошибки. Разбор схемы простой сети для распознавания цифр. Ограничения применения нейронных сетей.

Тема 8. Анализ текстовой информации и аналитика.

Основные методы добычи и анализа текстовых данных с целью обнаружения закономерностей. Постановка задачи. Области применения (маркетинг, социология, юриспруденция, медиа, HR). Основные задачи: классификация, кластеризация, суммаризация, извлечение сущностей, анализ тональности. Этапы анализа. Сбор и очистка данных (удаление шума, нормализация). Токенизация (слова, предложения, n-граммы). Лемматизация и стемминг. Удаление стоп-слов. Векторизация: Bag-of-Words, TF-IDF, word embeddings (Word2Vec, FastText). Методы и алгоритмы. Статистические подходы. Нейронные сети. LDA для тематического моделирования. NER (извлечение именованных сущностей). Алгоритмы классификации (логистическая регрессия, SVM, нейронные сети). Современные подходы. Трансформеры и BERT: принцип работы.

Python-библиотеки для анализа текстов: NLTK, spaCy, scikit-learn, gensim, transformers. API для анализа тональности (Яндекс, Google Cloud). Примеры решения практических задач: токенизация и лемматизация короткого текста, построение TF-IDF-матрицы для малого корпуса.

4.3 Практические занятия

Таблица 4

Содержание практических занятий и контрольные мероприятия

Название раздела, темы	№ и название лекций/ практических занятий	Формируемые компетенции (индикаторы)	Вид контрольного мероприятия	Кол-во Часов/ из них практическая подготовка
Тема 1. Основы науки о данных (Data Science).	Практическая работа 1. «Основы науки о данных (Data Science)».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Устный опрос, тест	2
Тема 2. Предобработка данных и их визуализация.	Практическая работа 2. «Предобработка данных и их визуализация.	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Защита работы, кейс, тест	4
Тема 3. Отбор признаков.	Практическая работа 3 «Отбор признаков».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Защита работы, тест	4
Тема 4. Обучение с учителем.	Практическая работа 4. «Обучение с учителем».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-4 (ПКос-4.1, ПКос-4.2, ПКос-4.3)	Защита работы, тест	4/2
Тема 5. Поиск ассоциативных правил в процессе анализа данных.	Практическая работа 5. «Поиск ассоциативных правил в процессе анализа данны».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Защита работы, кейс, тест	4
Тема 6. Кластерный анализ	Практическая работа 6. «Кластерный анализ».	ПКос-2 (ПКос-2.1, ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1,	Защита работы, тест	4/2

Название раздела, темы	№ и название лекций/ практических занятий	Формируемые компетенции (индикаторы)	Вид контрольного мероприятия	Кол-во Часов/ из них практическая подготовка
		ПКос-3.2, ПКос-3.3)		
Тема 7. Нейронные сети	Практическая работа 7. «Нейронные сети».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Защита работы, тест	4
Тема 8. Анализ текстовой информации и аналитика	Практическая работа 8. «Анализ текстовой информации и аналитика».	ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3) ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)	Защита работы, тест	4

Таблица 5

Перечень вопросов для самостоятельного изучения дисциплины

№ п/п	Название раздела, темы	Перечень рассматриваемых вопросов для самостоятельного изучения
1.	Тема 1. «Основы науки о данных (Data Science)».	История развития науки о данных. Современные тренды. ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3)
2.	Тема 2. «Предобработка данных и их визуализация».	Анализ полных наблюдений. Множественное восстановление пропущенных данных. Описание принципов качественной визуализации. Основные тенденции в области визуализации. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))
3.	Тема 3. «Отбор признаков».	Поиск взаимосвязей. Отбор факторов и снижение размерности исходных данных. Структура и шум в данных. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-4 (ПКос-4.1, ПКос-4.2, ПКос-4.3))
4.	Тема 4. «Обучение с учителем».	Анализ регрессионных остатков. Формальная и эффективная размерность. Графическая проверка линейности. Объясненная и необъясненная вариация. Оценка информативности признаков. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))
5.	Тема 5. «Поиск ассоциативных правил в процессе анализа данных».	Анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))
6.	Тема 6. «Кластерный анализ».	Итеративная оптимизация разбиения на кластеры. Плоские методы кластеризации. Метод ISODATA. Метод FOREL. Графовые методы. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))
7.	Тема 7. «Нейронные сети».	Компьютерное зрение. Современные архитектуры нейронных сетей. Анализ и синтез речи. Обработка естественного языка. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))
8.	Тема 8. «Анализ текстовой информации и аналитика».	Мешок слов. Классификация текстов. (ПКос-2 (ПКос-2.1, ПКос-2.2, ПКос-2.3), ПКос-3 (ПКос-3.1, ПКос-3.2, ПКос-3.3))

5. Образовательные технологии

Таблица 6

Применение активных и интерактивных образовательных технологий

№ п/п	Тема и форма занятия		Наименование используемых активных и интерактивных образовательных технологий (форм обучения)
1.	Тема 1. «Основы науки о данных (Data Science)».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы

№ п/п	Тема и форма занятия		Наименование используемых активных и интерактивных образовательных технологий (форм обучения)
2.	Тема 2. «Предобработка данных и их визуализация».	ПЗ	Мозговой штурм, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
3.	Тема 3. «Отбор признаков».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
4.	Тема 4. «Обучение с учителем».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
5.	Тема 5. «Поиск ассоциативных правил в процессе анализа данных».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
6.	Тема 6. «Кластерный анализ».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
7.	Тема 7. «Нейронные сети».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы
8.	Тема 8. «Анализ текстовой информации и аналитика».	ПЗ	Анализ конкретных ситуаций, интерактивные задания и тесты, мультимедийные презентации, видео и аудиоматериалы

6. Текущий контроль успеваемости и промежуточная аттестация по итогам освоения дисциплины

6.1. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений и навыков и (или) опыта деятельности

Контрольная работа №1

1. Загрузить данные в Python. Рассчитать по показателям основные статистики (среднюю, дисперсию, коэффициент вариации, медиану). Выбрать переменную с наибольшей вариацией признака. Проверить любые 2 переменные на соответствие нормальному закону. Сделать вывод.

2. Для целевой переменной провести классификацию методами:
- ближайший сосед;
- дерево решений;
- случайный лес;
- логистическая регрессия;
- опорные векторы.

При этом разделить выборки 70% - обучающая, 30% - тестовая.
Оценить качество классификации. Какому методу следует отдать предпочтение?

Примерное содержание практических работ

Практическая работа 1. «Основы науки о данных (Data Science)».

Цель: проверить понимание базовых понятий, типов задач, ключевых проблем и формальной постановки задачи машинного обучения.

1. История и общие понятия

Когда и в каком контексте зародилось наука о данных (DS)? Назовите 1–2 ключевых учёных/работ, повлиявших на становление DS.

Что такое машинное обучение? Сформулируйте определение своими словами и приведите один пример практической задачи.

В чём принципиальное отличие DS от классического программирования? Поясните на примере задачи распознавания спама.

Перечислите 3–4 области применения DS (кроме распознавания спама). Для каждой укажите одну конкретную задачу.

2. Типы задач и методы

Какие два основных типа задач обучения с учителем вы знаете? Кратко опишите каждую и приведите пример.

Что такое классификация? Назовите 2 алгоритма, решающих задачи классификации.

Что такое регрессия? Приведите пример задачи регрессии и 2 алгоритма для её решения.

Перечислите три типа задач машинного обучения (по наличию меток и механизму обучения). Для каждого типа укажите:

есть ли размеченные данные;

пример задачи.

Как классифицируют методы ML по принципу работы (например, линейные модели, ансамбли, нейронные сети)? Назовите по 1–2 метода для каждой группы.

3. Обобщающая способность и проблемы обучения

Что означает «обобщающая способность» (generalization) классификатора? Почему она важна?

Что такое переобучение (overfitting)? Приведите пример из реальной задачи. Как визуально распознать переобучение (например, по кривым обучения)?

Что такое недообучение (underfitting)? Какие признаки указывают на недообучение?

Перечислите 2–3 способа борьбы с переобучением.

Зачем делят данные на обучающую, валидационную и тестовую выборки? Какова роль каждой из них?

4. Формальная постановка задачи

Сформулируйте задачу машинного обучения в общем виде:

что является входными данными (X);

что — выходными (y);

что ищет алгоритм (гипотеза h);

что минимизируется (функция потерь L).

Что такое функция потерь (loss function)? Приведите 2 примера для классификации и 2 — для регрессии.

Что такое гипотеза (модель) в ML? Чем отличаются параметрические и непараметрические модели? Приведите по 1 примеру.

Что значит «обучить модель»? Какие параметры настраиваются в процессе обучения?

Почему важно иметь независимую тестовую выборку для оценки качества? Что такое «утечка данных» (data leakage) и как она искажает оценку?

Практическая работа 2. Предобработка данных и их визуализация.

Цель: отработать навыки очистки, преобразования и визуализации данных для последующего использования в моделях машинного обучения.

Задание 1. Загрузка и первичный анализ данных.
Загрузите датасет: titanic, iris, ... из sklearn; любой датасет из kaggle;
собственный CSV-файл (например, данные о продажах, клиентах и т.п.).

Выполните первичный анализ:
выведите первые 5 строк (head());
определите размерность (shape) и типы данных (dtypes);
проверьте наличие пропусков (isnull().sum());
для числовых столбцов выведите статистику (describe());
для категориальных — частоту значений (value_counts()).

Результат: отчёт с выводами о структуре и качестве данных (2–3 предложения).

Задание 2. Очистка данных
Удалите дубликаты (если есть):

Обработайте пропуски:
для числовых столбцов: заполните средним/медианой (fillna());
для категориальных: заполните наиболее частым значением или меткой "unknown".
Проверьте, что пропуски устранены.

Результат: очищенный датасет без дубликатов и пропусков.

Задание 3. Преобразование признаков
Кодируйте категориальные переменные:
примените OneHotEncoder (для признаков без порядка);
или LabelEncoder (для упорядоченных категорий).

Создайте новые признаки (опционально):
из даты — извлеките месяц/день недели (если есть столбец с датой);
объедините два категориальных признака в один (например, «город + район»).
Удалите ненужные столбцы (например, идентификаторы, если они не несут информации).

Результат: преобразованный датасет с числовыми и закодированными признаками.

Задание 4. Масштабирование и нормализация
Выделите числовые признаки (X_num).
Примените масштабирование:
StandardScaler (приведение к среднему = 0, std = 1);
MinMaxScaler (диапазон [0, 1]).
Сравните распределения до и после масштабирования (гистограммы для 1–2 признаков).

Результат: масштабированные числовые данные, графики сравнения распределений.

Задание 5. Визуализация данных
Постройте графики:
гистограммы для 2–3 числовых признаков (с подписями осей и заголовком);
коробчатые диаграммы (boxplot) для выявления выбросов;
тепловую карту корреляции (sns.heatmap(df.corr(), annot=True));
столбчатую диаграмму частоты категорий для 1–2 категориальных признаков.
Добавьте подписи и легенды к графикам.

Интерпретируйте визуализации:
какие признаки сильно коррелируют?
есть ли выбросы? Как их можно обработать?
какие категории встречаются чаще всего?
Результат: 4–5 графиков с подписями и краткий анализ (3–4 предложения).

Задание 6. Подготовка итогового датасета
Объедините масштабированные числовые и закодированные категориальные признаки.
Проверьте размерность итогового датасета (shape).
Сохраните результат в CSV.

Результат: файл processed_data.csv и вывод о готовности данных к моделированию.

Практическая работа 3. Отбор признаков.

Цель: освоить методы отбора наиболее информативных признаков для построения моделей машинного обучения; научиться оценивать влияние отбора признаков на качество модели.

Задание 1. Загрузка и первичный анализ данных
Загрузите датасет: библитечный датасет или собственный CSV-файл с признаками > 10 (например, данные о клиентах, продажах, технических параметрах).

Выполните первичный анализ:
выведите первые 5 строк (head());
определите размерность (shape) и типы данных (dtypes);
проверьте наличие пропусков (isnull().sum());
для числовых столбцов выведите статистику (describe());
постройте тепловую карту корреляции (sns.heatmap(df.corr(), annot=True)).
Результат: отчёт (2–3 предложения) о структуре данных и заметных корреляциях.

Задание 2. Предобработка данных
Обработайте пропуски (если есть):
числовые — заполните медианой;
категориальные — наиболее частым значением.
Кодируйте категориальные переменные (если есть) через OneHotEncoder или pd.get_dummies().
Масштабируйте числовые признаки с помощью StandardScaler.
Разделите данные на признаки (X) и целевую переменную (y).

Разбейте выборку на обучающую и тестовую (80 / 20 %).
Результат: подготовленные X_train, X_test, y_train, y_test.

Задание 3. Отбор признаков фильтрационными методами
Оцените важность признаков через корреляцию с целевой переменной:
для регрессии — коэффициент Пирсона;
для классификации — критерий хи-квадрат или взаимная информация.
Выберите топ-5–10 признаков с наибольшими значениями метрики.
Постройте бар-диаграмму важности признаков.
Результат: список отобранных признаков и график их значимости.

Задание 4. Отбор признаков оболочечными методами (wrapper)
Используйте метод рекурсивного исключения признаков (RFE) с моделью LogisticRegression или RandomForestClassifier.
Задайте желаемое число признаков (например, 5–7).
Выведите список отобранных признаков и их ранги.
Результат: список признаков, выбранных RFE, и объяснение их ранга.

Задание 5. Отбор признаков встроенными методами (embedded)
Обучите модель с L1-регуляризацией (Lasso для регрессии или LogisticRegression с penalty='l1' для классификации).
Выведите коэффициенты модели и отметьте, какие признаки получили нулевой вес (исключены).
Сравните список отобранных признаков с результатами из заданий 3–4.
Результат: список значимых признаков по L1-регуляризации и сравнение с предыдущими методами.

Задание 6. Оценка влияния отбора признаков на модель
Обучите одну и ту же модель (например, RandomForestClassifier):
на полном наборе признаков;
на отобранных признаках (по любому из методов выше).
Оцените качество на тестовой выборке:
для классификации — accuracy, F1, ROC-AUC;
для регрессии — MSE, RMSE, R^2.
Сравните метрики и время обучения. Постройте график сравнения (например, barplot для accuracy).
Результат: таблица метрик для двух версий модели и вывод о влиянии отбора признаков.

Задание 7. Интерпретация и выводы
Ответьте на вопросы:
Какие методы отбора дали наиболее согласованные результаты?
Какие признаки стабильно оказывались важными? Почему?
Как отбор признаков повлиял на интерпретируемость модели?
Можно ли ещё сократить число признаков без потери качества?
Предложите 1–2 способа дальнейшей оптимизации отбора (например, комбинация методов, подбор гиперпараметров).
Результат: 4–5 предложений с выводами и предложениями.

Практическая работа 4. Обучение с учителем.

Цель: освоить полный цикл построения модели обучения с учителем: от подготовки данных до оценки качества; научиться сравнивать алгоритмы и интерпретировать результаты.

Задание 1. Загрузка и анализ данных
Загрузите датасет:
вариант A: digits (классификация, из sklearn.datasets);
вариант B: california_housing (регрессия, из sklearn.datasets);
вариант C: CSV-файл с публичной задачей классификации/регрессии (например, Titanic, Boston Housing).

Выполните первичный анализ:
выведите первые 5 строк (head());
определите размерность (shape) и типы данных (dtypes);
проверьте пропуски (isnull().sum());
для числовых столбцов — статистику (describe());
постройте гистограммы целевых переменных;
оцените баланс классов (для классификации) или распределение целевой переменной (для регрессии).
Результат: отчёт (3–4 предложения) о структуре данных и потенциальных проблемах (пропуски, дисбаланс).

Задание 2. Предобработка данных
Обработайте пропуски:
числовые — медиана/среднее;
категориальные — наиболее частое значение.
Кодируйте категориальные переменные (OneHotEncoder или pd.get_dummies()).
Масштабируйте числовые признаки (StandardScaler или MinMaxScaler).

При дисбалансе классов (классификация) примените:
undersampling/oversampling (например, RandomUnderSampler, SMOTE);
веса классов в модели (class_weight='balanced').

Разделите данные на X_train, X_test, y_train, y_test (80 / 20 %).
Результат: готовые выборки для обучения и тестирования.

Задание 3. Обучение базовых моделей
Выберите 2–3 алгоритма для вашей задачи:
классификация: LogisticRegression, RandomForestClassifier, SVC, KNN;
регрессия: LinearRegression, DecisionTreeRegressor, GradientBoostingRegressor.
Обучите модели на X_train, y_train.
Сделайте прогнозы на X_test.
Результат: обученные модели и векторы предсказаний y_pred для каждой.

Задание 4. Оценка качества моделей
Для классификации выведите:
accuracy, precision, recall, F1 (макро/микро — по выбору);
матрицу ошибок (confusion_matrix);

ROC-AUC (для бинарной задачи) или мультиклассовый аналог.

Для регрессии выведите:

MSE, RMSE, MAE;
R^2;
график «истинные vs предсказанные».

Сравните модели в таблице:
Модель Метрика 1 Метрика 2 Время обучения
Результат: таблица сравнения, графики, вывод о лучшей модели.

Задание 5. Настройка гиперпараметров
Выберите одну модель из лучших по качеству.

Подберите гиперпараметры:
через GridSearchCV (полный перебор);
или RandomizedSearchCV (случайный поиск).
Используйте кросс-валидацию (5-fold).
Выведите лучшие параметры и оценку CV.
Переобучите модель на лучших параметрах и оцените на тестовой выборке.
Результат: лучшие гиперпараметры, сравнение качества до/после настройки.

Задание 6. Интерпретация модели
Для линейной модели (LogisticRegression / LinearRegression):
выведите коэффициенты (coef_) и свободный член (intercept_);
определите самые влиятельные признаки.

Для дерева/ансамбля:
постройте график важности признаков (feature_importances_).
Ответьте на вопросы:
Какие признаки сильнее всего влияют на прогноз?
Как интерпретировать знак коэффициента (для линейных моделей)?
Согласуется ли важность признаков с предметной логикой задачи?
Результат: график важности признаков, 3–4 предложения с интерпретацией.

Задание 7. Анализ ошибок
Для лучшей модели:
выделите объекты с наибольшими ошибками (например, самые «неверные» предсказания в классификации);
проанализируйте их характеристики (есть ли общие паттерны?).
Постройте кривую обучения (learning_curve) для диагностики переобучения/недообучения.
Предложите 1–2 способа улучшения модели (например, добавление признаков, смена алгоритма).
Результат: 3–5 предложений с выводами и предложениями.

Практическая работа 5. Поиск ассоциативных правил в процессе анализа данных.

Цель: освоить методику поиска ассоциативных правил на реальных данных; научиться настраивать параметры алгоритма, интерпретировать результаты и применять их для решения бизнес-задач.

Задание 1. Подготовка и загрузка данных
Выберите датасет:
вариант А: синтетическая транзакционная база (предоставляется преподавателем);

вариант B: публичный датасет «Market Basket Optimization» (Kaggle);
вариант C: CSV-файл с чеками магазина (столбцы: transaction_id, item).

Загрузите данные и выполните первичный анализ:
выведите первые 10 строк;
определите число уникальных транзакций и товаров;
постройте гистограмму распределения числа товаров в транзакции;
оцените долю частых/редких товаров (топ-10 и аутсайдеры).
Результат: отчёт (3–4 предложения) о структуре данных и особенностях.

Задание 2. Преобразование данных в формат «транзакционная база»
Приведите данные к формату «одна транзакция — одна строка»:
группировка по transaction_id, объединение товаров в список/строку;
либо создание бинарной матрицы (товар × транзакция) с флагами 0/1.
Удалите редкие товары (встречаются < 2 % транзакций) для ускорения расчётов.
Сохраните итоговый набор в формате, пригодном для алгоритма Apriori.
Результат: очищенная транзакционная база (DataFrame или sparse-матрица).

Задание 3. Поиск ассоциативных правил алгоритмом Apriori
Установите параметры алгоритма:
минимальная поддержка (min_support) — от 0,01 до 0,1 (подберите экспериментально);
минимальная достоверность (min_confidence) — от 0,2 до 0,6;
максимальная длина правила (max_len) — 2–4 товара.

Запустите Apriori и получите часто встречающиеся наборы товаров (frequent_itemsets).
Сгенерируйте ассоциативные правила на основе найденных наборов.
Выведите топ-10 правил по поддержке и топ-10 по достоверности.
Результат: таблицы правил с колонками:
antecedents (условие),
consequents (следствие),
support,
confidence,
lift.

Задание 4. Анализ и фильтрация правил
Отфильтруйте правила по критериям:
поддержка > 0,02;
достоверность > 0,3;
лифт > 1,2 (правила с лифтом ≤ 1 считайте неинформативными).
Выделите правила с высоким лифтом (> 2) — они указывают на сильную связь.
Найдите правила, где условие и следствие содержат по 1 товару (простые пары).
Постройте график распределения поддержки и достоверности для всех правил.
Результат: отфильтрованный набор правил и графики.

Задание 5. Интерпретация и визуализация
Для топ-5 правил ответьте на вопросы:
Какой товар чаще выступает условием, а какой — следствием?
Имеет ли правило практический смысл (например, «покупая хлеб, берут молоко»)?
Можно ли использовать правило для кросс-продаж или размещения товаров?

Постройте:
тепловую карту поддержки для пар товаров (top-15);

граф ассоциативных правил (узлы — товары, рёбра — правила, толщина/цвет — lift или confidence).

Результат: графики, скриншоты визуализаторов, 4–5 предложений с интерпретацией.

Задание 6. Практические выводы и рекомендации

Предложите 2–3 бизнес-решения на основе найденных правил:

какие товары разместить рядом;

какие связки предложить в акциях;

какие редкие товары добавить в кросс-продажи.

Оцените, как изменение параметров (min_support, min_confidence) влияет на число и качество правил.

Укажите ограничения вашего анализа (например, малый объём данных, отсутствие контекста покупок).

Результат: 5–7 предложений с рекомендациями и оценкой ограничений.

Практическая работа 6. Кластерный анализ.

Цель: освоить основные методы кластерного анализа; научиться подготавливать данные, подбирать параметры алгоритмов, оценивать качество кластеризации и интерпретировать результаты.

Задание 1. Загрузка и предварительный анализ данных

Выберите датасет:

вариант A: iris (из sklearn.datasets);

вариант B: CSV-файл с данными о клиентах/товарах/событиях (например, данные о покупках, характеристиках продуктов).

Выполните первичный анализ:

выведите первые 5 строк (head());

определите размерность (shape) и типы данных (dtypes);

проверьте пропуски (isnull().sum());

для числовых столбцов — статистику (describe());

постройте парные диаграммы рассеяния (pairplot) для визуализации взаимосвязей;

оцените масштаб признаков и необходимость нормализации.

Результат: отчёт (3–4 предложения) о структуре данных, выявленных проблемах и планах предобработки.

Задание 2. Предобработка данных

Обработайте пропуски (если есть):

числовые — заполните медианой/средним;

категориальные — наиболее частым значением.

Кодируйте категориальные переменные (при наличии) через OneHotEncoder или pd.get_dummies().

Масштабируйте числовые признаки (StandardScaler или MinMaxScaler).

При необходимости удалите выбросы (например, через Z-оценку или IQR).

Сформируйте матрицу признаков X для кластеризации.

Результат: очищенная и нормализованная матрица признаков X.

Задание 3. Кластеризация методом K-means

Определите диапазон числа кластеров k (например, от 2 до 10).

Для каждого k:

обучите модель KMeans;
сохраните инерцию (inertia_) и коэффициент силуэта (silhouette_score).
Постройте график «локтя» (inertia vs k) и график силуэта vs k.
Выберите оптимальное k на основе графиков.
Обучите финальную модель K-means с оптимальным k, получите метки кластеров.

Визуализируйте результаты:
для 2D/3D-данных — точечный график с цветом по кластерам;
для многомерных данных — PCA-проекция + раскраска по кластерам.
Результат: графики для выбора k, финальная визуализация кластеров, метки labels_kmeans.

Задание 4. Иерархическая кластеризация
Постройте матрицу расстояний между объектами (евклидово расстояние).
Примените агломеративную кластеризацию (AgglomerativeClustering) с разными метриками связи (ward, complete, average).
Постройте дендрограмму для визуализации иерархии кластеров.
Выберите число кластеров, соответствующее разрезу дендрограммы.
Сравните метки кластеров с результатами K-means (например, через adjusted_rand_score).
Визуализируйте кластеры (аналогично п. 3.6).
Результат: дендрограмма, визуализация кластеров, метки labels_hierarchical, сравнение с K-means.

Задание 5. Кластеризация DBSCAN
Подберите параметры:
eps (радиус окрестности) — через анализ распределения расстояний до ближайших соседей;
min_samples (минимальное число точек в окрестности) — от 3 до 10.
Обучите модель DBSCAN, получите метки кластеров (включая шум, метка −1).
Оцените число кластеров и долю шумовых точек.
Визуализируйте результаты (как в п. 3.6), выделив шум отдельным цветом.
Сравните с K-means и иерархической кластеризацией.
Результат: визуализация DBSCAN, метки labels_dbscan, анализ шума и параметров.

Задание 6. Оценка качества кластеризации
Для каждого метода (K-means, иерархический, DBSCAN) рассчитайте:
коэффициент силуэта (silhouette_score);
индекс Калинского-Харабаша (calinski_harabasz_score);
индекс Дэвиса-Болдина (davies_bouldin_score).

Составьте таблицу сравнения:

Метод	Силуэт	Калинский-Харабаш	Дэвис-Болдин	Число кластеров
K-means	…	…	…	…
Иерархический	…	…	…	…
DBSCAN	…	…	…	…

Сделайте вывод: какой метод дал наилучшее разделение?
Результат: таблица метрик, 2–3 предложения с выводами.

Задание 7. Интерпретация кластеров
Для лучшего метода (по метрикам из п. 6) проанализируйте кластеры:
выведите средние значения признаков для каждого кластера (groupby(labels).mean());

найдите признаки, максимально различающие кластеры;
дайте названия/описания кластерам на основе их характеристик.

Ответьте на вопросы:
Какие объекты попали в один кластер? Есть ли логичная общая черта?
Как можно использовать кластеры в бизнес-задаче (например, сегментация клиентов)?
Есть ли аномальные кластеры (очень малые или с высокой дисперсией)?
Результат: таблица средних по кластерам, 4–5 предложений с интерпретацией и предложениями.

Практическая работа 7. Нейронные сети.

Цель: освоить основы построения и обучения нейронных сетей; научиться подготавливать данные, выбирать архитектуру, обучать модель, оценивать качество и интерпретировать результаты.

Задание 1. Реализуйте класс для обучения и работы простейшей нейронной сети.

Задание 2. Решить задачу регрессии с использованием стандартного датасета (например, Boston Housing).

Задание 3. Решить задачу классификации с использованием стандартного датасета (например, Titanic).

Задание 4. Решить задачу распознавания рукописных цифр с использованием датасета MNIST.

Задание 5. Постройте графики:
потери (loss) на обучении и валидации;
точности (accuracy) на обучении и валидации.

Оцените признаки переобучения/недообучения.
Результат: графики обучения, вывод о динамике качества.

Задание 6. Оценка качества модели
Сделайте предсказания на тестовой выборке (model.predict()).

Для классификации рассчитайте:
точность (accuracy);
матрицу ошибок (confusion_matrix);
precision, recall, F1 (по классам и макро/микро);
ROC-AUC (для бинарной задачи).

Для регрессии рассчитайте:
MSE, RMSE, MAE;
R^2;
график «истинные vs предсказанные».

Сравните качество с базовыми моделями (например, логистическая регрессия, случайный лес).
Результат: таблица метрик, графики, сравнение с альтернативными моделями.

Задание 7. Настройка гиперпараметров
Подберите оптимальные параметры через:
GridSearchCV / RandomizedSearchCV (для простых моделей);
инструменты Keras Tuner / Optuna (для нейросетей).

Исследуйте:
число слоёв и нейронов;
скорость обучения (learning_rate);
функции активации;
регуляризацию (Dropout, L1/L2).
Переобучите модель с лучшими параметрами и оцените качество на тесте.
Сравните результаты до и после настройки.
Результат: лучшие гиперпараметры, таблица метрик до/после настройки.

Ответьте на вопросы:
Какие признаки/паттерны модель считает важными?
Как архитектура влияет на качество?
Каковы ограничения модели?
Результат: визуализации фильтров/активаций, 4–5 предложений с интерпретацией.

Задание 8. Анализ ошибок
Выделите объекты с наибольшими ошибками (неверные классы, высокие остатки).
Проанализируйте их характеристики:
есть ли общие паттерны (например, сложные изображения, выбросы)?
относятся ли они к определённому классу/группе?
Предложите способы улучшения:
увеличение объёма данных;
аугментация (для изображений);
изменение архитектуры.

Результат: 3–4 предложения с выводами и предложениями.

Практическая работа 8. Анализ текстовой информации и аналитика.

Цель: освоить базовые методы обработки и анализа текстовых данных; научиться извлекать смысловые паттерны, проводить количественную и качественную оценку текстов, применять инструменты текстовой аналитики для решения прикладных задач.

Задание 1. Загрузить текстовые данные, выполнить предобработку, подсчитать базовые статистики. (набор отзывов (CSV/JSON); новостные статьи (папка с .txt-файлами); социальные медиа (твиты, посты)).

Задание 2. Применить методы частотного и коллокационного анализа, визуализировать результаты.

Задание 3. Определить тональность текста.

Задание 4. Тематический анализ и интерпретация (выявить темы в корпусе, оценить качество кластеризации текстов, сформулировать выводы).

Вопросы для подготовки к устным опросам
Тема 1. Основы науки о данных (Data Science).

1. Особенности наукки о данных.
2. Примеры применения на практике науки о данных
3. Применение методов data science в сельском хозяйстве

4. Современные тренды развития науки о данных.

Тема 2. Предобработка данных и их визуализация.

1. Характеристики инструментов визуализации данных.
2. Методы визуализации.
3. Существующие тренды в области визуализации данных.
4. 3D визуализация.

Тема 3. Отбор признаков.

1. Что такое отбор признаков (feature selection) и зачем он нужен в машинном обучении?

2. В чём отличие отбора признаков от извлечения признаков (feature extraction)?

3. Перечислите 3–4 основных преимущества отбора признаков для модели и процесса обучения.

4. Какие проблемы решает отбор признаков (переобучение, интерпретируемость, скорость и т. п.)?

5. Что такое «шумные» (нерелевантные) признаки? Приведите пример из реальной задачи.

6. Что подразумевается под «избыточными» признаками? Как они влияют на модель?

Тема 5. Обучение с учителем.

1. Что означает термин «обучение с учителем»? В чём его ключевое отличие от обучения без учителя и обучения с подкреплением?

2. Какие данные необходимы для обучения с учителем? Опишите структуру обучающей выборки (признаки и метки).

3. Перечислите 3–4 типичные задачи, решаемые методами обучения с учителем. Приведите по одному реальному примеру для каждой задачи.

4. Что такое размеченные данные? Почему их подготовка — трудоёмкий этап в supervised learning?

5. В чём суть процесса обучения модели с учителем? Как модель «учится» на примерах?

6. Что такое функция потерь (loss function)? Приведите 2 примера функций потерь для классификации и регрессии.

7. Объясните понятия:
обучающая выборка;
валидационная выборка;
тестовая выборка.
Какова роль каждой из них?

Тема 4. Поиск ассоциативных правил в процессе анализа данных.

1. Что такое ассоциативные правила (association rules)? Сформулируйте цель их поиска в данных.

2. Приведите 2–3 реальных примера использования ассоциативных правил (в ритейле, веб-аналитике, медицине и т. п.).

3. Что такое «рыночная корзина» (market basket) в контексте анализа ассоциативных правил?

4. Какие типы данных обычно анализируют при поиске ассоциативных правил (категориальные, количественные, последовательности)?

5. В чём отличие поиска ассоциативных правил от поиска последовательных шаблонов?

Тема 6. Кластерный анализ.

1. Что такое кластерный анализ? В чём его ключевое отличие от классификации с учителем?

2. Сформулируйте основные цели кластерного анализа. Приведите 3–4 примера прикладных задач (в маркетинге, биологии, социологии и т. п.).

3. Что означает «однородность» внутри кластера и «различие» между кластерами? Как эти понятия формализуются?

4. Какие типы данных можно анализировать с помощью кластерного анализа (числовые, категориальные, смешанные)?

5. Что такое мера расстояния (сходства) в кластерном анализе? Назовите 2–3 популярные метрики (например, евклидово расстояние, манхэттенское).

6. В чём особенность нормализации данных перед кластерным анализом? Когда она обязательна?

7. Опишите общий алгоритм кластеризации k-средних (k-means). Какие шаги он включает?

8. Как выбрать число кластеров k? Перечислите 2–3 метода (например, метод локтя, силуэтный анализ).

9. Что показывает силуэтный коэффициент (silhouette score)? Как его интерпретировать?

10. В чём суть иерархической кластеризации? Чем отличаются агломеративные и дивизивные методы?

11. Что такое дендрограмма? Как по ней определить оптимальное число кластеров?

12. Опишите методы связи в иерархической кластеризации: одиночная связь (ближайший сосед); полная связь (дальний сосед); средняя связь. В чём их особенности?

13. Что такое DBSCAN? Какие параметры задаются в этом алгоритме (eps, min_samples)? Как они влияют на результат?

14. В чём преимущество DBSCAN перед k-means? Для каких данных он особенно полезен?

Тема 7. Нейронные сети.

1. Что такое искусственная нейронная сеть (ИНС)? В чём её сходство и отличие от биологической нейронной сети?

2. Опишите структуру искусственного нейрона: входные данные, веса, сумматор, функция активации, выход.

3. Какие функции активации вы знаете? Приведите 3–4 примера (например, сигмоида, ReLU, tanh) и кратко опишите их свойства.

4. Что такое «веса» нейронов и как они настраиваются в процессе обучения?

5. В чём суть прямого распространения сигнала (forward pass) в нейронной сети?

6. Что такое полносвязная (fully-connected) нейронная сеть? Как устроены её слои?

7. Чем отличается входной, скрытый и выходной слой сети? Какова их роль?

8. Что означает термин «глубокое обучение» (deep learning)? В чём особенность глубоких сетей?

9. Что такое обучающая выборка для нейронной сети? Какие данные она должна содержать?

10. В чём заключается обучение с учителем для нейронных сетей? Приведите пример задачи.

11. Что такое функция потерь (loss function) в нейронных сетях? Приведите 2 примера для классификации и регрессии.

12. Как работает алгоритм обратного распространения ошибки (backpropagation)? Опишите основные шаги.

13. Что такое градиентный спуск? В чём отличие SGD, Adam, RMSprop?

14. Что такое эпоха (epoch) и батч (batch) в обучении нейронных сетей?

15. Что такое переобучение (overfitting) в нейронных сетях? Назовите 2–3 способа борьбы с ним.

16. Что такое регуляризация в нейронных сетях? Сравните L1 и L2-регуляризацию.

17. Что такое дропаут (dropout)? Как он помогает предотвратить переобучение?

18. Опишите архитектуру свёрточной нейронной сети (CNN). Для каких задач она применяется?

Тема 8. Анализ текстовой информации и аналитика

1. Основные этапы текстового анализа.
2. Задачи текстового анализа.
3. Извлечение ключевых понятий из текста.
4. Классификация документов.
5. Кластерный анализ документов.
6. Существующие программные обеспечения в области анализа текстовой информации.

Кейсы

Кейс 1. «Основные термины и понятия открытых данных».

Цель: через практическую ситуацию закрепить термины «открытые данные», «набор данных», «машиночитаемый формат», «открытая лицензия», «персональные данные», «анонимизация» и др.

Проблемная ситуация (АПК).

Региональное министерство сельского хозяйства запускает портал открытых данных АПК. Планируется публиковать:

- реестр сельскохозяйственных производителей (с указанием вида деятельности, площади земель, типов продукции);
- статистику урожайности по районам и культурам за последние 10 лет;
- данные о применении удобрений и пестицидов по хозяйствам в агрегированном виде;
- сведения о господдержке (субсидии по направлениям, без раскрытия ФИО фермеров).

Студенты должны предложить идеи ИИ-сервисов (прогноз урожайности, оценка рисков, рекомендации по посевным структурам) и определить, какие из перечисленных наборов могут быть опубликованы как открытые данные и в каком виде.

Задания студентам.

1. Идентификация наборов открытых данных АПК

Студенты выбирают, какие наборы могут стать открытыми при соблюдении условий:

- реестр производителей: опубликовать в виде открытого набора, если это юрлица/ИП и указаны только общедоступные реквизиты (ИНН/ОГРН, адрес хозяйства, виды продукции), без персональных данных физических лиц;
- статистика урожайности: может быть открытой, если данные агрегированы по району/культуре и не позволяют идентифицировать конкретное фермерское хозяйство;

- применение удобрений и пестицидов: публикация в агрегированном виде по району, типу культуры и классу вещества;

- господдержка: суммы по программам и категориям получателей в разрезе районов без указания ФИО фермеров.

Студенты указывают:

- какие из этих наборов явно являются «наборами открытых данных», если размещены в машиночитаемом формате (CSV, JSON) и снабжены открытой лицензией;

- какие наборы требуют анонимизации или агрегирования для исключения утечки персональных данных.

2. Персональные данные и анонимизация в АПК

Студенты определяют, где потенциально возникают персональные данные:

- фермеры-физлица (ФИО, адрес личного хозяйства, паспортные данные, персональные телефоны);

- данные о господдержке, если можно однозначно связать сумму с конкретным фермером.

Необходимо предложить способы анонимизации:

- замена ФИО на код/идентификатор или публикация только агрегированных данных по группам;

- укрупнение географии (район вместо конкретного населенного пункта), если малая численность получателей может привести к повторной идентификации.

3. Форматы и лицензии

Студенты формулируют рекомендации:

- в каких машиночитаемых форматах публиковать данные АПК (CSV для табличных данных, GeoJSON/SHAPE для пространственных данных, JSON/XML для API-доступа);

- какие условия должна содержать открытая лицензия (право на свободное использование, переработку и распространение при указании источника, отсутствие ограничений по сфере использования).

Отдельно обсуждается, почему без явной открытой лицензии даже технически доступные данные АПК нельзя считать полноценно «открытыми».

Результаты работы.

По итогам работы студенты:

- заполняют таблицу с колонками: «Набор данных АПК», «Можно ли публиковать как открытые данные?», «Требуемая анонимизация/агрегация», «Рекомендуемый формат», «Лицензионные условия»;

- формулируют своими словами термины «открытые данные», «набор открытых данных АПК», «машиночитаемый формат», «открытая лицензия», «персональные данные», «анонимизация» применительно к сельскому хозяйству.

Кейс 2. «Построение рядов распределения в Excel, Python. Описательные статистики в Excel, Python»

Анализ выборки объектов недвижимости: студенты строят ряды распределения и вычисляют характеристики выборки в Excel и Python для ключевых количественных признаков (цены, площади, количества комнат).

Проблемная ситуация.

Есть датасет квартир в крупном городе (например, выгрузка с портала объявлений):

- price — цена, тыс. руб.;
- area — общая площадь, м²;
- rooms — число комнат;
- floor, floors_total, district и др.

Цель: подготовить аналитическую записку для агентства недвижимости:

- построить ряды распределения по количеству комнат и по цене;
- вычислить среднюю, моду и медиану цены и площади;
- сравнить результаты Excel и Python и интерпретировать различия.

Задачи.

1. Ряды распределения в Python

Студенты:

- загружают датасет в pandas;
- строят негруппированный ряд распределения по rooms (частоты для 1, 2, 3, 4+ комнат);
- строят интервальный ряд по price (например, 0–3000, 3000–6000 тыс. руб. и т.д.).
- добавляют относительные частоты (деление частот на объем выборки).

По area студенты также формируют интервалы (до 30 м², 30–50, 50–70, 70+ м²), чтобы увидеть структуру предложения по площади.

2. Ряд распределения и статистики в Excel

Студенты:

- выгружают те же данные (или подвыборку) в Excel;

- строят частотную таблицу по rooms (считать частоты по уникальным значениям);

- формируют интервальный ряд по price (либо вручную, либо через гистограмму/«Анализ данных»);

- считают:

  - среднюю цену (функция среднего);

  - медиану цены;

  - моду цены (или модальное значение по интервалам).

Аналогично — средняя, медиана и (при необходимости) мода по площади area.

3. Описательные статистики в Python

Студенты:

- вычисляют mean, median, mode для price и area через методы pandas или функции numpy/statistics;

- убеждаются, что результаты совпадают (с точностью до округления) с Excel, при условии одинаковой выборки и учета пропусков;

- строят простую визуализацию (гистограмму цены или площади) для наглядного сравнения с таблицей распределения.

Обсуждаются эффекты: вытянутость распределения цены вправо из-за дорогих объектов, положение медианы относительно средней, возможная многомодальность по количеству комнат (например, пики у 1- и 2-комнатных).

Ожидаемый результат: итоговый мини-отчет.

В отчете студенты представляют:

- таблицу ряда распределения по rooms и по интервалам price (частоты и доли);

- значения средней, медианы и моды цены и площади (с указанием, в какой среде посчитаны — Excel/Python);

- короткий текстовый вывод:

  - какая мера (средняя или медиана) лучше описывает «типичную» цену квартиры при наличии дорогих выбросов;

  - как структура по количеству комнат отражает профиль рынка (например, доминирование 1–2-комнатных квартир).

Кейс 3. «Прогнозирование цен недвижимости»

Цель: применение линейной регрессии для построения прогностической модели.

Проблемная ситуация. Вы работаете аналитиком в агентстве недвижимости. Руководству необходимо предсказать цены квартир на основе площади и количества комнат.

Задачи:

1. Загрузить данные (площадь, количество комнат, цена).
2. Разделить их на обучающую и тестовую выборки.
3. Реализовать простую линейную регрессию с использованием библиотеки scikit-learn.
4. Построить графики зависимости фактических и предсказанных значений с помощью matplotlib или plotly.
5. Рассчитать метрики ошибки (MAE, MSE, R^2).

Ожидаемый результат: обучающая программа, прогнозирующая стоимость квартиры по введённым характеристикам.

Кейс 4. «Классификация студентов по успеваемости»

Цель: освоение базовых принципов классификации и метрик качества модели.

Проблемная ситуация. Отдел учебной аналитики университета хочет определить, какие студенты находятся в группе риска по результатам успеваемости.

Задачи:

1. Использовать набор данных о студентах (оценки, посещаемость, выполнение домашних заданий).
2. Реализовать бинарную классификацию с использованием алгоритма логистической регрессии или дерева решений.
3. Разделить данные на обучающую и тестовую выборки.
4. Построить confusion matrix и вычислить метрики точности (accuracy, precision, recall, F1-score).

Ожидаемый результат: модель, определяющая вероятность попадания студента в группу риска.

Кейс 5. «Сегментация клиентов интернет-магазина»

Цель: изучение методов кластеризации без учителя.

Проблемная ситуация. Интернет-магазин хочет разделить клиентов на группы по поведению (средний чек, частота покупок, время последней активности).

Задачи:

1. Подготовить и нормализовать данные клиентов.
2. Реализовать метод K-Means и подобрать оптимальное число кластеров (метод локтя).
3. Визуализировать результаты с использованием двумерного графика кластеров.
4. Описать поведенческие особенности каждой найденной группы клиентов.

Ожидаемый результат: выявленные кластеры клиентов и их характеристика для маркетингового анализа.

Кейс 6. «Классификация методом машинного обучения «Дерево решений»»

Цель: кейс ориентирован на практическое знакомство с классификацией методом дерева решений: постановка задачи, подготовка данных, обучение модели и интерпретация правил «если… то…» на реальном примере.

Проблемная ситуация.

Компания по аренде недвижимости хочет автоматически классифицировать заявки клиентов на два класса:

- «высокий риск отказа» (клиент, скорее всего, откажется от сделки или не пройдет скоринг);
- «низкий риск отказа» (клиент с высокой вероятностью завершит сделку).

Есть датасет заявок с признаками: возраст, уровень дохода, наличие просрочек по кредитам, длительность аренды в прошлом, число несовершеннолетних детей, тип занятости и целевая метка `target` (0 — низкий риск, 1 — высокий риск).

Задачи:

1. Постановка задачи и подготовка данных

- Определить тип задачи: бинарная классификация методом дерева решений.
- Разделить выборку на обучающую и тестовую (например, 70/30).
- Обсудить, какие признаки могут влиять на риск отказа и нужно ли их кодировать (категориальные признаки, пропуски).

2. Обучение дерева решений

- Обучить модель дерева решений на обучающей выборке (в Python с использованием библиотеки машинного обучения).
- Настроить базовые гиперпараметры: максимальная глубина, минимальное число объектов в листе, критерий качества (Gini/entropy).
- Оценить качество на тестовой выборке с помощью доли верных ответов и, при необходимости, матрицы ошибок.

3. Интерпретация дерева и правил

Студентам нужно:

- визуализировать дерево или вывести текстовое представление основных ветвей;
- извлечь несколько человекочитаемых правил вида:
  - если доход < порога и есть просрочки, то «высокий риск»;
  - если доход ≥ порога и нет просрочек, то «низкий риск»;
- объяснить, почему дерево выбрало именно такие признаки и пороги в верхних узлах (наиболее информативные признаки).

4. Анализ ограничений и улучшений

Обсудить:

- риск переобучения при слишком глубоком дереве;
- влияние дисбаланса классов (если «высокий риск» встречается редко);
- идеи по улучшению: ограничение глубины, минимальный размер листа, использование ансамблей (случайный лес, градиентный бустинг) как развитие темы.

Ожидаемый результат:

Студенты готовят краткий отчет:

- описание исходной задачи и признаков;
- метрики качества модели на тесте (accuracy, при желании precision/recall);
- 3–5 ключевых правил дерева в человекочитаемом виде и их интерпретация для бизнеса;
- вывод о пригодности дерева решений для автоматической предварительной оценки заявок и о рисках его использования.

Кейс 7. «Нейронные сети».

Цель: базовое знакомство с нейронными сетями: от постановки задачи и подготовки данных до обучения простой модели и интерпретации результатов в практическом сценарии.

Проблемная ситуация.

Финтех-стартап хочет предсказывать вероятность отклика клиента на предложение по новой банковской карте. У компании есть датасет исторических маркетинговых кампаний с признаками: возраст, доход, количество продуктов в банке, наличие мобильного банка, количество обращений в колл-центр, участие в прошлых акциях и целевая метка `responded` (1 — откликнулся, 0 — нет).

Задача для студентов: построить простую нейронную сеть для бинарной классификации (`responded`) и сравнить ее работу с логистической регрессией или решающим деревом.

Задачи:

1. Постановка задачи и подготовка данных.

Студенты:

- формулируют задачу как задачу бинарной классификации по табличным данным;
- делят выборку на обучающую и тестовую (например, 70/30);
- выполняют предобработку:
  - нормализацию числовых признаков (масштабирование);
  - кодирование категориальных признаков (one-hot);

- проверку и обработку пропусков.

Обсуждается, почему нейросети чувствительны к масштабу признаков и корректному кодированию категориальных переменных.

2. Архитектура простой нейросети.

Студенты проектируют базовую архитектуру:

- входной слой: размерность = число признаков после кодирования;
- 1–2 скрытых слоя (например, 16–32 нейрона) с нелинейной активацией (ReLU);
- выходной слой из 1 нейрона с сигмоидой для предсказания вероятности отклика.

Выбираются:

- функция потерь — бинарная кросс-энтропия;
- оптимизатор — простой градиентный метод (например, Adam);
- метрика качества — доля верных ответов, дополнительно по возможности F1 или AUC.

3. Обучение и оценка модели.

Студенты:

- обучают нейронную сеть на обучающей выборке с мини-батчами и несколькими эпохами;
- фиксируют динамику функции потерь и качества на обучающей и валидационной выборках (overfitting/underfitting);
- оценивают качество на тестовой выборке, сравнивая результаты с более простой моделью (логистическая регрессия или дерево решений).

Обсуждается:

- даёт ли нейросеть выигрыш по качеству по сравнению с простой моделью;
- как ограничение числа слоёв и нейронов влияет на переобучение и время обучения.

4. Интерпретация и практический вывод.

Студенты:

- анализируют, какие группы клиентов чаще всего получают предсказание «откликнется» (например, возраст 25–40, активный мобильный банк, 2–3 продукта);
- строят простые визуализации (распределение предсказанных вероятностей, ROC-кривую при наличии инструментария);
- формируют рекомендации для маркетинга:
  - какие сегменты целевой аудитории приоритетны;
  - на каких признаках модель, судя по анализу, «фокусируется» (по косвенным признакам, важности признаков из простой модели и т.п.).

Ожидаемый результат:

От группы ожидается мини-отчёт, включающий:

- постановку задачи и описание признакового пространства;
- архитектуру нейронной сети (слои, активации, размерности) и использованные гиперпараметры;
- сравнительную таблицу метрик для нейросети и базовой модели;
- интерпретацию результатов на языке предметной области (для руководства стартапа, а не для ML-специалистов).

Примерные вопросы к зачету

1. Основы понятия науки о данных (Data Science).
2. Основные этапы разработки моделей машинного обучения.
3. Сбор данных для реализации модели машинного обучения.
4. Разведочный анализ данных.
5. Линейная регрессия.
6. Метод наименьших квадратов (МНК).
7. Метод градиентного спуска.
8. Метрики и метрические пространства.
9. Алгоритм KNN.
10. Деревья решений.
11. Ансамбли моделей. Случайный лес
12. Метод опорных векторов
13. Метод опорных векторов с ядерной функцией
14. Бустинг
15. Бэггинг
17. Градиентный бустинг
18. Логистическая регрессия
19. Дискриминантный анализ
20. Метрики качества моделей машинного обучения
21. Наивный байесовский метод.
22. ID3 и C4.5 алгоритмы для построения деревьев решений.
23. Ассоциативные правила.
24. Априорный алгоритм.
25. Кластерный анализ. Классификация алгоритмов кластеризации.
26. Иерархические алгоритмы: агломерационные и дивизионные методы кластеризации.
27. Неиерархические алгоритмы: k-means, Fuzzy C-Means.
28. Основные этапы анализа текста.
29. Задачи добычи текста.
30. Извлечение центральных понятий из текста.
31. Классификация текстовых документов.
32. Кластеризация текстовых документов.
33. Пакеты программ для анализа текста.

34. Характеристика средств визуализации данных. Методы визуализации.
35. Основные тенденции в визуализации данных.

Тесты

1. Задача классификации – это _____ задача.
a) описательная
b) предсказательная
c) качественная
d) количественная

2. Задача кластеризации – это _____ задача.
a) описательная
b) предсказательная
c) качественная
d) количественная

3. Единицы наблюдения, которые значительно отличаются от большинства других единиц в наборе данных:
a) транзакция
b) порядковое число
c) интервалы
d) резко выделяющиеся значения

4. Набор конкретных примеров с известным исходом:
a) клиент-сервер
b) классификатор
c) учебный набор
d) интеллектуальный анализ данных

5. Преобразование данных включает в себя…
a) разделение данных из одного источника на несколько источников данных
b) процесс изменения данных с обобщенного уровня на более детализированный
c) объединение данных из одного источника с другими источниками данных
d) процесс изменения данных с детализированного уровня на более обобщенный

6. _____ - полезный метод обнаружения закономерностей в начале процесса интеллектуального анализа данных.
a) расчёт меры расстояния
b) дерево принятия решений
c) ассоциативные правила
d) приёмы визуализации

7. _____ - класс моделей, принцип которых основан на аналогии с работой человеческого мозга:
a) нейронные сети
b) кластеры
c) дерево принятия решений
d) правило классификации

8. Какая иерархическая структура у деревьев принятия решений?
a) ЕСЛИ… ТО…
b) НИ… НИ…
c) ЛИБО… ЛИБО…
d) КАК… ТАК И…

9. Что из перечисленного является математическим уравнением, связывающим переменные x и y?
a) регрессия
16b) интерполяция
c) кластеризация
d) экстраполяция

10. Уравнение вида y = a + bx:
a) полиномиальное уравнение
b) линейная регрессия
c) регрессия
d) интерполяция

11. Какой из этих показателей является измерением качества модели регрессии?
a) средняя арифметическая
b) дисперсия
c) стандартное отклонение
d) коэффициент детерминации

12. Коэффициент корреляции всегда лежит между значениями…
a) 0 и 1
b) -1 и 1
c) -1 и 0
d) 0 и 2

13. Изучение взаимосвязи между несколькими переменными – это задача…
a) парной регрессии
b) множественной регрессии
c) дерева принятия решения
d) моделирования

14. Что из перечисленного является методом построения правил классификации?
a) 1R-алгоритм
b) метод Naive Bayes
c) оба варианта верны
d) ни один из вариантов не является верным

15. Ожидаемое значение y, когда X = 0, равно:
a) коэффициенту полной регрессии
b) условному началу
c) коэффициенту корреляции
d) коэффициенту детерминации

16. Собранные в разные моменты времени значения каких-либо параметров - это…
a) панельные данные
b) пространственные данные
c) данные временного ряда
d) ни один из вариантов

17. Изучение взаимосвязи благодаря…
a) двумерному графику
b) гистограммам
c) графикам временных рядов
d) ни один из вариантов

18. Главная цель поиска ассоциативных правил заключается в том, чтобы…
a) создать правила классификации
b) проверить достоверность регрессионной модели
c) определить главную идею той или иной текстовой информации
d) выявить закономерности между связанными событиями в базах данных

19. Если несколько событий связаны друг с другом, то это…
a) ассоциация
b) последовательность
c) классификация
d) кластеризация

20. Отношение транзакций, которые имеют набор F (DF) к общему количеству транзакций (D) называется…
a) весомым уровнем набора F
b) моделью набора F

c) кластером набора F
d) уровнем поддержки набора F

21. Набор предметов (F) называется частым, когда…
a) Supp (F) < Supp (min)
b) Supp (F) > Supp (min)
c) Supp (F) = Supp (min)
d) ни один из вариантов

22. Объекты некоторого набора предметов, которые подвергаются анализу, называются…
a) пропорции
b) центроиды
c) транзакции
d) модели

23. Что является последовательностью в поиске ассоциативных правил?
a) конечные действительные числа
b) ранжированный ряд
c) упорядоченное множество некоторых объектов
d) ни один из вариантов

24. Что из перечисленного не является оценкой полезности ассоциативных правил?
a) качество
b) поддержка
c) достоверность
d) улучшение

25. Поддержка любого набора объектов не может превышать минимальной поддержки любого из его подмножеств. Это основное свойство…
a) метода Naive Bayes
b) алгоритма Apriori
c) дерева принятия решений
d) кластеризации

26. Один из методов кластерного анализа называется…
a) стандартное отклонение
b) k-средние
c) регрессия
d) дисперсия

27. Какой из алгоритмов является иерархическим?
a) агломеративный

b) метод k-средних
c) метод нечеткой кластеризации C-средних
d) ни один из вариантов

28. Что из перечисленного не является методом пересчёта расстояний между кластерами?
a) расстояние между ближайшими соседями
b) расстояние между дальними соседями
c) метод медиан
d) модальный метод

29. Какой из алгоритмов является неиерархическим?
a) агломеративный
b) дивизимный
c) метод k-средних
d) ни один из вариантов

30. Что такое кластеры?
a) данные временных рядов
b) однородные группы объектов
c) разнородные группы объектов
d) статистическая совокупность

31. Одной из мер близости, используемой в кластеризации, является…
a) расстояние Чебышева
b) Евклидово расстояние
c) оба варианта верны
d) ни один из вариантов не является верным

32. Общепринятый способ визуализации результатов кластерного анализа является построение…
a) гистограммы
b) двумерного графика
c) дендрограммы
d) графика временного ряда

33. Что такое качество кластеризации?
a) степень приближения результата кластеризации к идеальному решению
b) правильный выбор количества кластеров
c) мера идеального расстояния между кластерами
d) пригодность полученных результатов для дальнейшего исследования

34. Что из перечисленного не является подходом к оценке поисковых информационных систем?
a) полнота (recall)

b) выпадение (fall-out)
c) закономерности (patterns)
d) точность (precision)

35. Что является первым этапом в анализе текстовой информации?
a) предварительная обработка документов
b) извлечение информации из текста
c) интерпретация результатов
d) поиск информации

36. Что такое стемминг?
a) морфологический поиск
b) поисковая система
c) текстовый документ
d) интересная закономерность

37. Классификация документов является синонимом к слову…
a) стемминг
b) аннотирование
c) категоризация
d) ни один из вариантов

38. Одной из задач анализа текстовой информации является…
a) кластеризация
b) извлечение ключевых понятий
c) классификация
d) всё вышеперечисленное

39. Что является примером текстовых данных?
a) веб-страницы
b) e-mail
c) нормативные документы
d) всё вышеперечисленное

40. Слова, которые являются вспомогательными и несут
мало информации о содержании документа, - это…
a) N-граммы
b) стоп-слова
c) диалекты
d) слова из веб-страниц
и

41. Отношением числа релевантных документов, информационно-поисковой системой, к общему числу
документов называется…

a) точностью
b) полнотой
c) условием
d) эффектом

42. График может быть удобным представлением данных, если…
a) существует взаимосвязь между объектами данных
b) объекты данных показывают определенную тенденцию
c) оба варианта верны
d) ни один из вариантов не является верным

43. Первым этапом визуализации данных является…
a) беглый обзор
b) построение графика
c) интерпретация результатов
d) оценка эффективности

44. Одним из методов визуализации является…
a) кластеризация
b) категоризация
c) геометрические преобразования
d) ни один из вариантов

45. Что из перечисленного не является графиком?
a) гистограмма
b) круговая диаграмма
c) ранжированный ряд
d) лепестковая диаграмма

46. Визуализация данных позволяет нам обнаружить…
a) закономерности
b) тренды
c) корреляции
d) всё вышеперечисленное

6.2. Описание показателей и критериев контроля успеваемости, описание шкал оценивания

В седьмом семестре для оценки знаний, умений, навыков и формирования компетенции по дисциплине может применяться **балльно-рейтинговая** система контроля и оценки успеваемости студентов.

В основу балльно-рейтинговой системы (БРС) положены принципы, в соответствии с которыми формирование рейтинга студента осуществляется в ходе текущего контроля и промежуточной аттестации знаний.

Максимальное количество баллов, которое может набрать студент за работу в семестре - 200 баллов. Из них 100 баллов - текущая работа, 40 - премиальные баллы (могут быть начислены за активную работу на занятиях, выполнение дополнительных заданий, участие в профильных мероприятиях), 60 баллов - промежуточная аттестация. Студент допускается к сдаче зачета при достижении рейтинга 60 баллов.

Оценка знаний студента формируется как сумма баллов за текущую работу и промежуточную аттестацию.

Текущая работа состоит из выполнении и защит практических работ, прохождении тестов, выполнении кейсов.

Тест (0-10 баллов).

Защита работ (0-64 баллов). Каждая выполненная и сданная работа оценивается от 0 до 8 баллов.

8 баллов - ставится при наличии незначительных неточностей в ответе.

6 балла - при наличии негрубых ошибок в ответе, которые не привели к ложным выводам и неверному пониманию сути вопроса.

4 балла - сделаны неверные выводы по применяемым методам, при этом общее понимание применяемых методов не искажено.

0-2 балла - нарушена логика в понимании применяемых методов.

Устный опрос (тема 1) оценивается от 0 до 6 баллов.

Кейсы оцениваются от 0 до 10 баллов. (2 обязательных кейса - 20 баллов).

Участие в интерактивных занятиях может быть зачтено активным студентам как участие в опросе по теме, на котором применялись интерактивные технологии.

На зачете студент может получить максимальное количество баллов равное 60. Далее итоговая оценка определяется следующим образом.

Промежуточный контроль в пятом семестре – экзамен.

Таблица 7

Шкала оценивания (средний балл)	**Экзамен**
> 140	Отлично
120-140	Хорошо
101-119	Удовлетворительно
0-100	Неудовлетворительно

Положительными оценками, при получении которых дисциплина засчитывается в качестве пройденной, являются оценки «удовлетворительно», «хорошо» и «отлично».

Если получена оценка «неудовлетворительно» по дисциплине, то необходимо, после консультации с преподавателем, в течение 10 календарных дней следующего семестра подготовить ответы на ряд вопросов, предусмотренных программой обучения, и представить результаты этих ответов преподавателю.

Критерии оценивания результатов обучения

Таблица 8

Оценка	Критерии оценивания
Высокий уровень «5» (отлично)	оценку **«отлично»** заслуживает студент, освоивший знания, умения, компетенции и теоретический материал без пробелов; выполнивший все задания, предусмотренные учебным планом на высоком качественном уровне; практические навыки профессионального применения освоенных знаний сформированы. Компетенции, закреплённые за дисциплиной, сформированы на уровне – высокий.
Средний уровень «4» (хорошо)	оценку **«хорошо»** заслуживает студент, практически полностью освоивший знания, умения, компетенции и теоретический материал, учебные задания не оценены максимальным числом баллов, в основном сформировал практические навыки. Компетенции, закреплённые за дисциплиной, сформированы на уровне – хороший (средний).
Пороговый уровень «3» (удовлетворительно)	оценку **«удовлетворительно»** заслуживает студент, частично с пробелами освоивший знания, умения, компетенции и теоретический материал, многие учебные задания либо не выполнил, либо они оценены числом баллов близким к минимальному, некоторые практические навыки не сформированы. Компетенции, закреплённые за дисциплиной, сформированы на уровне – достаточный.
Минимальный уровень «2» (неудовлетворительно)	оценку «неудовлетворительно» заслуживает студент, не освоивший знания, умения, компетенции и теоретический материал, учебные задания не выполнил, практические навыки не сформированы. Компетенции, закреплённые за дисциплиной, не сформированы.

7. Учебно-методическое и информационное обеспечение дисциплины

7.1 Основная литература

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2025. — 89 с. — (Высшее образование). — ISBN 978-5-534-20732-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/558662 (дата обращения: 07.08.2025).
2. Воронина, В. В. Теория и практика машинного обучения : учебное пособие / В. В. Воронина. — Ульяновск : УлГТУ, 2017. — 290 с. — ISBN 978-5-9795-1712-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165053.
3. Демичев, В. В. Алгоритмизация и программирование: Учебное пособие / В. В. Демичев, Д. В. Быков, Д. Э. Храмов [и др.].; рец. С.Г. Сальников; Российский государственный аграрный университет - МСХА имени К.А. Тимирязева (Москва). — Электрон. текстовые дан. — Москва, 2024. — 248 с. — Коллекция: Учебная и учебно-методическая литература. — Свободный доступ из сетиИнтернет (чтение, печать, копирование). — Режим доступа : http://elib.timacad.ru/dl/full/s17122024AP_v3.pdf. - Загл. с титул. экрана.

-Электрон. версия печ. публикации. — <URL:http://elib.timacad.ru/dl/full/s17122024AP_v3.pdf>.
4. Бессмертный, И. А. Интеллектуальные системы : учебник и практикум для вузов / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2025. — 250 с. — (Высшее образование). — ISBN 978-5-534-20734-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/558664 (дата обращения: 01.06.2025).
5. Федоров, Д. Ю. Программирование на python : учебное пособие для вузов / Д. Ю. Федоров. — 6-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 187 с. — (Высшее образование). — ISBN 978-5-534-19666-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/556864 (дата обращения: 15.08.2025).7.2

Дополнительная литература

1.Зыков, С. В. Программирование : учебник и практикум для вузов / С. В. Зыков. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 285 с. — (Высшее образование). — ISBN 978-5-534-16031-4. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/560815 (дата обращения: 10.08.2025).

2.Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики : пер. с англ. / под ред. Ю. В. Кирютенко. — 2-е изд. — М. : Вильямс, 2009. — 784 с. — ISBN 978-5-8459-1588-6.

3.Чернышев, С. А. Основы программирования на Python : учебник для вузов / С. А. Чернышев. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 349 с. — (Высшее образование). — ISBN 978-5-534-17139-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/567821 (дата обращения: 17.08.2025).

4.Кормен, Т. Х. Алгоритмы: построение и анализ / Томас Х. Кормен, Чарльз Э. Лейзерсон, Рональд Л. Ривест, Клиффорд Штайн; пер. с англ. под ред. И. В. Красикова. — 2-е изд. — М. : Вильямс, 2012. — 1296 с. — ISBN 978-5-8459-1794-2.

5.Скиена, С. С. Алгоритмы. Руководство по разработке / Стивен С. Скиена ; пер. с англ. А. Л. Семёнова; под ред. А. К. Звонкова. — 2-е изд. — М.: ДМК Пресс, 2011. — 720 с. — ISBN 978-5-94074-714-0.

6.Стивенс, Род. Алгоритмы. Теория и практическое применение: [численные алгоритмы, структуры данных, методы работы с массивами, связанными списками и сетями] / Род Стивенс; [пер.: Кириленко Вадим, Волошко Роман Влади-

мирович]. — Москва: Э, 2016. — 542, с. : ил., табл.; 24 см. — (Мировой компьютерный бестселлер). — ISBN 978-5-699-81729-0.
7.Бхаргава А. Грокаем алгоритмы. Иллюстрированное пособие для программистов и любопытствующих = Grokking Algorithms / пер. с англ. А. В. Белова. — Санкт-Петербург: Питер, 2017. — 288 с. : ил. — ISBN 978-5-496-02513-9.
8.Станкевич, Л. А. Интеллектуальные системы и технологии : учебник и практикум для вузов / Л. А. Станкевич. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 478 с. — (Высшее образование). — ISBN 978-5-534-20363-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/560754 (дата обращения: 01.07.2025).
9.Кудрявцев, В. Б. Интеллектуальные системы : учебник и практикум для вузов / В. Б. Кудрявцев, Э. Э. Гасанов, А. С. Подколзин. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2025. — 165 с. — (Высшее образование). — ISBN 978-5-534-07779-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/561954 (дата обращения: 02.08.2025).
10.Navarro G., Nekrich Y. Top-k Document Retrieval in Compressed Space // Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). – Philadelphia: SIAM, 2025. – P. 4009–4030. – DOI: 10.1137/1.9781611978322.137.
11.Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting AI-generated text: Factors influencing detectability with current methods. Journal of Artificial Intelligence Research, 82, 2233–2278. https://doi.org/10.1613/jair.1.16665
12.Cheng S.-W., Huang H. Fréchet Distance in Subquadratic Time // Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). – Philadelphia : SIAM, 2025. – P. 5100–5113. – DOI: 10.1137/1.9781611978322.173.
13.Ellert J., Gawrychowski P., Górkiewicz A., Starikovskaya T. Faster two-dimensional pattern matching with k mismatches // Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). – Philadelphia : SIAM, 2025. – P. 4031–4060. – DOI: 10.1137/1.9781611978322.138.

7.3 Статьи, опубликованные в научных журналах 1 уровня Белого списка научных журналов Минобрнауки России и сборниках научных работ конференций уровня А*

1. Picon, A., Eguskiza, I., Galan, P., Gomez-Zamanillo, L., Romero, J., Klukas, C., Bereciartua-Perez, A., Scharner, M., & Navarra-Mestre, R. (2025). Crop-conditional semantic segmentation for efficient agricultural disease assessment. Artificial Intelligence in Agriculture, 15(1), 79–87. https://doi.org/10.1016/j.aiia.2025.01.002.

2. Mittal, S., Thakral, K., Singh, R. et al. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. Nat Mach Intell 6, 936–949 (2024). https://doi.org/10.1038/s42256-024-00874-y/.

3. Alistarh D., Kurtic E., Malinovsky G., Modoranu I.-V., Richtárik P., Robert T., Safaryan M. MicroAdam: Accurate Adaptive Optimization with Low Space Complexity // Advances in Neural Information Processing Systems 37: Proc. of the 37th Conf. on Neural Information Processing Systems (NeurIPS 2024, Vancouver, Canada, 10–15 Dec. 2024). – Neural Information Processing Systems Foundation, Inc., 2024. – P. 1–43. – DOI: 10.52202/079017-0001.

4. Kang M., Park Y., Song C. SPO: Sequential Monte Carlo Policy Optimisation // Advances in Neural Information Processing Systems 37: Proc. of the 37th Conf. on Neural Information Processing Systems (NeurIPS 2024, Vancouver, Canada, 10–15 Dec. 2024). – Neural Information Processing Systems Foundation, Inc., 2024. – P. 73–131. – DOI: 10.52202/079017-0003.

5. Гарбук С. В. Метод декомпозиции функциональных характеристик систем искусственного интеллекта // Искусственный интеллект и принятие решений. – 2025. – № 1. – С. 14–32. – DOI: 10.14357/20718594250102.

6. Yan J., Zhang W.-G., Liu Y., Pan W., Hou X.-Y., Liu Z.-Y. An autonomous navigation method for field phenotyping robot based on ground-air collaboration // Artificial Intelligence in Agriculture. – 2025. – Vol. 15, No. 4. – P. 610–621. – DOI: 10.1016/j.aiia.2025.05.005.Yang Y., Wang X., Zhang F., Wu Z., Wang Y., Wang J. MSNet: A multispectral-image driven rapeseed canopy instance segmentation network // Artificial Intelligence in Agriculture. – 2025. – Vol. 15, No. 4. – P. 642–658. – DOI: 10.1016/j.aiia.2025.05.008.

7.4 Методические указания, рекомендации и другие материалы к занятиям

1. Кудрина, Е. В. Основы алгоритмизации и программирования на языке C# : учебное пособие для вузов / Е. В. Кудрина, М. В. Огнева. — Москва : Издательство Юрайт, 2022. — 322 с. — (Высшее образование). — ISBN 978-5-534-09796-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/517285 (дата обращения: 18.08.2022).

2. Кочегурова, Е. А. Теория и методы оптимизации : учебное пособие для вузов / Е. А. Кочегурова. — Москва : Издательство Юрайт, 2022. — 133 с. — (Высшее образование). — ISBN 978-5-534-10090-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/490136 (дата обращения: 18.08.2022).

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Официальный сайт Python. URL: https://www.python.org/.
2. Официальный сайт дистрибутива языков программирования Python и R Anaconda. URL: https://www.anaconda.com/.
3. Международное сообщество разработчиков моделей машинного обучения KAGGLE. https://www.kaggle.com/.
4. Сообщество open data science. https://ods.ai/.
5. Официальный сайт Росстата. URL: https://rosstat.gov.ru/ .
6. Портал открытых данных Российской Федерации (https://data.gov.ru).
7. Портал открытых данных Правительства Москвы (https://data.mos.ru).
8. Каталог каталогов открытых данных (https://www.datacatalogs.ru).
9. Минфин России (https://minfin.gov.ru/ru/opendata/).
10. Единая точка доступа к открытым данным учреждений ЕС European data portal (https://data.europa.eu).
11. Open data portals Европейской комиссии (https://digital-strategy.ec.europa.eu/en/policies/open-data-portals).
12. Портал открытых данных правительства США Data.gov (https://data.gov).
13. Всемирный каталог открытых данных DataPortals.org (https://dataportals.org).
14. База данных о 2600+ порталах открытых данных в мире Open Data Inception (https://opendatainception.io).
15. Подборка международных открытых данных Open data resources (UK Data Service) (https://ukdataservice.ac.uk/help/other-data-providers/open-data-resources/).

9. Перечень программного обеспечения и информационных справочных систем

Таблица 9

Перечень программного обеспечения

№ п/п	Наименование раздела учебной дисциплины	Наименование программы	Тип программы	Автор	Год разработки
1	Тема 1. «Основы науки о данных (Data Science)».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023 /2007/2012 и позднее

2	Тема 2. «Предобработка данных и их визуализация».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023 /2007/2012 и позднее
3	Тема 3. «Отбор признаков».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023 /2007/2012 и позднее
4	Тема 4. «Обучение с учителем.	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023 /2007/2012 и позднее
5	Тема 5. «Поиск ассоциативных правил в процессе анализа данных».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023 /2007/2012 и позднее
6	Тема 6. «Кластерный анализ».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023/2007/2012 и позднее

7	Тема 7. «Нейронные сети».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023/2007/2012 и позднее
8	Тема 8. «Анализ текстовой информации и аналитика».	VS Code/LaTeX /Excel/Word/Anaconda (или свободно-распространяемые аналоги)	Кроссплатформенный текстовый редактор для разработчиков /Издательская система/ Редактор электронных таблиц/ Текстовый процессор/ Система управления пакетами и дистрибутив	Microsoft/LPPL/ Microsoft/Anaconda Inc. (или opensource)	2025/2023/2007/2012 и позднее

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

Таблица 10

Сведения об обеспеченности специализированными аудиториями, кабинетами, лабораториями

Наименование специальных помещений и помещений для самостоятельной работы (№ учебного корпуса, № аудитории)	Оснащенность специальных помещений и помещений для самостоятельной работы
1	2
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (2й учебный корпус, 102 ауд.)	Количество рабочих мест: 16 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE. Структурное подразделение: Институт Экономики и управления, Кафедра Статистики и кибернетики
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проек-	Корпус 2, Аудитория 106 Количество рабочих мест: 16 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость

тирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (2й учебный корпус, 106 ауд.)	10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE. Структурное подразделение: Институт Экономики и управления, Кафедра Статистики и кибернетики
учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации, помещение для самостоятельной работы (2й учебный корпус, 302 ауд.)	Корпус 2, Аудитория 302 Количество рабочих мест: 16 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE. Структурное подразделение: Институт Экономики управления, Кафедра Статистики и кибернетики
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (1й учебный корпус, 212 ауд.)	Корпус 1, Аудитория 212 Количество рабочих мест: 24 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE.Структурное подразделение: Кафедра Цифровая кафедра
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего кон-	Корпус 1, Аудитория 214 Количество рабочих мест: 24 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4

троля и промежуточной аттестации (1й учебный корпус, 214 ауд.)	ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE.Структурное подразделение: Кафедра Цифровая кафедра
Центральная научная библиотека имени Н.И. Железнова	Читальные залы библиотеки
Студенческое общежитие	Комната для самоподготовки

11. Методические рекомендации студентам по освоению дисциплины

Приступая к изучению дисциплины «Основы науки о данных (Data Science)», студенты должны ознакомиться с учебной программой, учебной, научной и методической литературой, имеющейся в библиотеке РГАУ-МСХА им. К.А. Тимирязева, получить в библиотеке рекомендованные учебники и учебно-методические пособия, завести новую тетрадь для работы с первоисточниками.

В ходе занятий вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации. Задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, разрешения спорных ситуаций.

В ходе подготовки к практическим занятиям изучить основную литературу, ознакомиться с дополнительной литературой в соответствии с поставленной задачей. При этом учесть рекомендации преподавателя и требования учебной программы. Необходимо дорабатывать свой конспект, делая в нем соответствующие записи из литературы, рекомендованной преподавателем и предусмотренной учебной программой.

При подготовке к экзамену (в конце семестра) повторять пройденный материал в строгом соответствии с учебной программой. Использовать конспекты и литературу, рекомендованную преподавателем. Обратить особое внимание на темы учебных занятий, пропущенных студентом по разным причинам. При необходимости обратиться за консультацией и методической помощью к преподавателю.

Виды и формы отработки пропущенных занятий

Студент, пропустивший занятия, обязан выполнить и защитить практические работы по теме пропущенных занятий. В рамках часов консультаций студент может сдать и защитить практические работы.

12. Методические рекомендации преподавателям по организации обучения по дисциплине

Курс «Основы науки о данных (Data Science)» должен давать не абстрактно-формальные, а прикладные знания. Данная цель может быть реализована только при условии соблюдения в учебных планах преемственности учебных дисциплин. Базовые знания для изучения «Основы науки о данных (Data

Science)» дают такие дисциплины, как «Специальные главы математики», «Модели информационных процессов и систем», «Статистика (продвинутый уровень)», «Эконометрика (продвинутый уровень)», «Инструменты Data Science в R, Python, SQL». Изучение основных тем данной дисциплины позволит студентам сформировать представление о предмете «Основы науки о данных (Data Science)», получить практические навыки решения основных оптимизационных задач и необходимые знания для последующего профессионального развития в этой области.

Студент может подготовить доклад по теме, представляющей его научный интерес, представить результаты в виде презентации. В случае надлежащего качества, его работа может быть заслушана на научном кружке кафедры или на студенческой научной конференции. По решению кафедры, студенты, занявшие призовые места на научных студенческих конференциях, могут освобождаться от сдачи экзамена по этой дисциплине.

Преподаватель должен указывать, в какой последовательности следует изучать материал дисциплины, обращать внимание на особенности изучения отдельных тем и разделов, помогать отбирать наиболее важные и необходимые сведения из учебных пособий, а также давать объяснения вопросам программы курса, которые обычно вызывают затруднения. При этом преподавателю необходимо учитывать следующие моменты:

1. Не следует перегружать студентов творческими заданиями.
2. Чередовать творческую работу на занятиях с заданиями во внеаудиторное время.
3. Давать студентам четкий инструктаж по выполнению самостоятельных заданий: цель задания; условия выполнения; объем; сроки; требования к оформлению.
4. Осуществлять текущий учет и контроль за самостоятельной работой.
5. Давать оценку и обобщать уровень усвоения навыков самостоятельной, творческой работы.

Программу разработал(и):

Калитвин В.А., канд. ф.-м. наук, доцент _____
(подпись)

РЕЦЕНЗИЯ
на рабочую программу дисциплины Б1.В.02 «Основы науки о данных (Data Science)» ОПОП ВО по направлению 09.04.02 Информационные системы и технологии, направленность «Науки о данных» (квалификация выпускника – магистр)

Прудким Александром Сергеевичем, доцентом кафедры высшей математики, кандида-том педагогических наук (далее по тексту рецензент), проведено рецензирование рабочей программы дисциплины «Основы науки о данных (Data Science)» ОПОП ВО по направлению 09.04.02 Информационные системы и технологии**,** направленность «Науки о данных» (магистратура) разработанной в ФГБОУ ВО «Российский государственный аграрный университет – МСХА имени К.А. Тимирязева», на кафедре статистики и кибернетики (разработчик – Калитвин Владимир Анатольевич, доцент кафедры статистики и кибернетики).

Рассмотрев представленные на рецензирование материалы, рецензент пришел к следующим выводам:

1. Предъявленная рабочая программа дисциплины «Основы науки о данных (Data Science)» (далее по тексту Программа) *соответствует* требованиям ФГОС ВО по направлению 09.03.02 Информационные системы и технологии. Программа *содержит* все основные разделы, *соответствует* требованиям к нормативно-методическим документам.

2. Представленная в Программе ***актуальность*** учебной дисциплины в рамках реализации ОПОП ВО *не подлежит сомнению* – дисциплина относится к дисциплинам части, формируемой участниками образовательных отношений учебного плана по направлению подготовки 09.04.02 Информационные системы и технологии – Б1.В.

3. Представленные в Программе ***цели*** дисциплины *соответствуют* требованиям ФГОС ВО направления 09.04.02 Информационные системы и технологии.

4. В соответствии с Программой за дисциплиной «Основы науки о данных (Data Science)» закреплены 2 ***профессиональных компетенции***. Дисциплина «Основы науки о данных (Data Science)» и представленная Программа *способна реализовать* их в объявленных требованиях.

5. ***Результаты обучения***, представленные в Программе в категориях знать, уметь, владеть *соответствуют* специфике и содержанию дисциплины и *демонстрируют возможность* получения заявленных результатов.

6. Общая трудоёмкость дисциплины «Основы науки о данных (Data Science)» составляет 2 зачётные единицы (72 часа/ из них практическая подготовка 4 ч.).

7. Информация о взаимосвязи изучаемых дисциплин и вопросам исключения дублирования в содержании дисциплин *соответствует* действительности. Дисциплина «Основы науки о данных (Data Science)» взаимосвязана с другими дисциплинами ОПОП ВО и Учебного плана по направлению *09.04.02 Информационные системы и технологии* и возможность дублирования в содержании отсутствует.

8. Представленная Программа предполагает использование современных образовательных технологий, используемые при реализации различных видов учебной работы. Формы образовательных технологий *соответствуют* специфике дисциплины.

9. Программа дисциплины «Основы науки о данных (Data Science)» предполагает проведение занятий в интерактивной форме.

10. Виды, содержание и трудоёмкость самостоятельной работы студентов, представленные в Программе, *соответствуют* требованиям к подготовке выпускников, содержащимся во ФГОС ВО направления 09.04.02 Информационные системы и технологии.

11. Представленные и описанные в Программе формы *текущей* оценки знаний (устный опрос, защита практических работ), *соответствуют* специфике дисциплины и требованиям к выпускникам.

Форма промежуточного контроля знаний студентов, предусмотренная Программой, осуществляется в форме зачета с оценкой в седьмом семестре, что *соответствует* статусу

дисциплины, как дисциплины части, формируемой участниками образовательных отношений учебного плана по направлению подготовки 09.04.02 Информационные системы и технологии – Б1.В ФГОС ВО направления 09.04.02.Информационные системы и технологии**.**

12. Формы оценки знаний, представленные в Программе, *соответствуют* специфике дисциплины и требованиям к выпускникам.

13. Учебно-методическое обеспечение дисциплины представлено: основной литературой – 5 источников (базовые учебники), дополнительной литературой – 15 наименований, статьи, опубликованные в научных журналах 1 уровня Белого списка научных журналов Минобрнауки России и сборниках научных работ конференций уровня А* - 6 шт., Интернет-ресурсы – 15 источников и *соответствует* требованиям ФГОС ВО направления *09.04.02 Информационные системы и технологии***.**

14. Материально-техническое обеспечение дисциплины соответствует специфике дисциплины «Методы машинного обучения» и обеспечивает использование современных образовательных, в том числе интерактивных методов обучения.

15. Методические рекомендации студентам и методические рекомендации преподавателям по организации обучения по дисциплине дают представление о специфике обучения по дисциплине «Основы науки о данных (Data Science)».

ОБЩИЕ ВЫВОДЫ

На основании проведенного рецензирования можно сделать заключение, что характер, структура и содержание рабочей программы дисциплины «Основы науки о данных (Data Science)» ОПОП ВО по направлению 09.04.02 Информационные системы и технологии, направленность «Науки о данных» (квалификация выпускника – магистр), разработанная Калитвиным Владимиром Анатольевичем, доцентом, кандидатом физико-математических наук, соответствует требованиям ФГОС ВО, современным требованиям экономики, рынка труда и позволит при её реализации успешно обеспечить формирование заявленных компетенций.

Рецензент: Прудкий А.С., доцент кафедры высшей математики ФГБОУ ВО «Российский государственный аграрный университет – МСХА имени К.А. Тимирязева», кандидат педагогических наук

«26» августа 2025 г.