

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Астраханский государственный университет имени В. Н. Татищева»
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО
Руководитель ОПОП

А.В. Синельщиков

«05» мая 2025 г.

УТВЕРЖДАЮ
И.о. зав. кафедрой
информационных технологий

О.Н. Выборнова

«05» мая 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Интеллектуальный анализ данных и машинное обучение

Составитель(и) **Синельщиков А.В., доцент, к.т.н., доцент кафедры ИТ**
Согласовано с работодателями: **Проталинский И.О., к.т.н., технический директор**
ООО «Бест Плюс»
Кузнецова В.Ю., к.т.н., финансовый директор ООО
«Сильвер код»

Направление подготовки / **09.04.02 ИНФОРМАЦИОННЫЕ СИСТЕМЫ И**
специальность **ТЕХНОЛОГИИ**

Направленность (профиль) / **ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ**
специализация ОПОП **ТЕХНОЛОГИИ**

Квалификация (степень) **магистр**

Форма обучения **очная**

Год приёма **2025**

Курс **2**

Семестр(ы) **3**

Астрахань – 2025 г.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целями освоения дисциплины (модуля) «Интеллектуальный анализ данных и машинное обучение» является формирование теоретических знаний и практических навыков в области применения методов интеллектуального анализа данных и машинного обучения для решения прикладных задач в информационных системах и технологиях, развитие компетенций, позволяющих анализировать большие объемы данных, строить прогнозные модели, выявлять закономерности и принимать обоснованные решения в условиях цифровой трансформации организаций..

1.2. Задачи освоения дисциплины (модуля) «Интеллектуальный анализ данных и машинное обучение»:

- Изучение теоретических основ интеллектуального анализа данных и машинного обучения.
- Ознакомление с основными алгоритмами и методами обработки данных.
- Формирование навыков подготовки и очистки данных для анализа и моделирования.
- Освоение техник построения и оценки моделей машинного обучения.
- Развитие умений применения инструментов и библиотек для реализации алгоритмов (например, Python, R).
- Подготовка к решению практических задач анализа данных в различных предметных областях.
- Выработка способности интерпретировать результаты моделей и использовать их для поддержки принятия решений.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) «Интеллектуальный анализ данных и машинное обучение» относится к элективным дисциплинам и осваивается во 3 семестре.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения и навыки, формируемые предшествующими дисциплинами (модулями):

- Системы поддержки принятия решений;
- Базы и хранилища данных;
- Модели информационных процессов.
- Математические основы искусственного интеллекта.

Знать:

- Основы высшей математики: ключевые понятия и методы линейной алгебры (векторы, матрицы, операции над ними), математического анализа (производные, градиенты), теории вероятностей и математической статистики (распределения, статистические тесты, оценка гипотез).
- Основные концепции и архитектуры систем поддержки принятия решений (СППР), методы извлечения и представления знаний.
- Ключевые понятия в области управления данными: модели данных, принципы реляционных баз данных, основы языка SQL, назначение и архитектура хранилищ данных.
- Основы моделирования информационных процессов, основные нотации (например, DFD, ERD) и базовые модели жизненного цикла ИС.

Уметь:

- Применять математический аппарат для формализации и решения базовых задач (например, использовать методы статистики для описания выборки, находить градиент функции).
- Формулировать запросы (SQL) для извлечения, фильтрации и агрегации данных из баз

данных.

- Интерпретировать (читать) модели информационных процессов и схемы данных (ER-диаграммы).
- Определять типы задач, решаемых с помощью СППР, и выбирать подходящие методы их решения на концептуальном уровне.

Владеть:

- Навыками использования математических и статистических методов для предварительного анализа данных.
- Методами концептуального проектирования баз данных и описания потоков данных.
- Навыками работы с данными на уровне извлечения и подготовки (с использованием SQL).
- Терминологией в области математической статистики, баз данных и систем поддержки принятия решений.

2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения, навыки, формируемые данной учебной дисциплиной (модулем):

- Облачные технологии
- Безопасность информационных систем
- Выпускная квалификационная работа.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс изучения дисциплины направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки (специальности):

- профессиональные:

ПК-3. Способен выполнять разработку программного обеспечения общего и прикладного назначения, проводить оценку качества и работоспособности разработанного программного обеспечения, осуществлять его сопровождение и модификацию.

ПК-5. Способен вести сдачу проекта, собирать и анализировать мнения и замечания заказчика по выполнению проекта и предлагать соответствующие решения

Таблица 1. Декомпозиция результатов обучения

Код и наименование компетенции	Планируемые результаты обучения по дисциплине (модулю)		
	Знать (1)	Уметь (2)	Владеть (3)

ПК-3. Способен выполнять разработку программного обеспечения общего и прикладного назначения, проводить оценку качества и работоспособности разработанного программного обеспечения, осуществлять его сопровождение и модификацию	ПК.3.1. Знать современные тенденции развития информационных технологий.	ПК.3.2. Уметь решать задачи по разработке программного обеспечения общего и прикладного назначения, проведения оценки качества и работоспособности разработанного программного обеспечения, осуществлять его сопровождение и модификацию.	ПК.3.3. Владеть навыками выбора и применения современных инструментальных средств для решения задач по разработке программного обеспечения общего и прикладного назначения, проведения оценки качества и работоспособности разработанного программного обеспечения, осуществления его сопровождения и модификации.
ПК-5. Способен вести сдачу проекта, собирать и анализировать мнения и замечания заказчика по выполнению проекта и предлагать соответствующие решения	ПК-5.1. Знать инструменты и методы управления выпуском и поставкой проектов в области ИТ.	ПК-5.2. Уметь проводить анализ мнений и замечаний заказчиков по выполнению проекта.	ПК-5.3. Владеть навыками формулирования решений по внесению изменений в ИТ-проекты по согласованию с заказчиками

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Общая трудоемкость дисциплины в соответствии с учебным планом составляет 4 зачетные единицы (144 часа).

Трудоемкость отдельных видов учебной работы студентов очной формы обучения приведена в таблице 2.1.

Таблица 2.1. Трудоемкость отдельных видов учебной работы по формам обучения

Вид учебной и внеучебной работы	для очной формы обучения
Объем дисциплины в зачетных единицах	4
Объем дисциплины в академических часах	144
Контактная работа обучающихся с преподавателем (всего), в том числе (час.):	54
- занятия лекционного типа, в том числе:	18
- практическая подготовка (если предусмотрена)	0
- занятия семинарского типа (семинары, практические, лабораторные), в том числе:	36
- практическая подготовка (если предусмотрена)	0
Самостоятельная работа обучающихся (час.)	90
Форма промежуточной аттестации обучающегося (зачет/экзамен), семестр (ы)	зачет с оценкой – 3 семестр

Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий и самостоятельной работы, для очной формы обучения представлено в таблице 2.2.

Таблица 2.2. Структура и содержание дисциплины (модуля) для очной формы обучения

Раздел, тема дисциплины (модуля)	Контактная работа, час.						КР/КП	СР, час.	Итого часов	Форма текущего контроля успеваемости, форма промежуточной аттестации
	Л		ПЗ		ЛР					
	Л	в т.ч. ПП	ПЗ	в т.ч. ПП	ЛР	в т.ч. ПП				
Тема 1. Введение в Data Science и методология KDD (Knowledge Discovery in Databases)	2				4			14	20	Собеседование. Отчет по лабораторной работы
Тема 2. Подготовка и предварительный анализ данных (Data Preparation & EDA).	2				4			14	20	Отчет по лабораторной работы
Тема 3. Инженерия признаков (Feature Engineering).	2				4			14	20	Отчет по лабораторной работы
Тема 4. Обучение с учителем (Supervised Learning).	4				8			16	28	Отчет по лабораторной работы
Тема 5. Обучение без учителя (Unsupervised Learning).	4				8			16	28	Отчет по лабораторной работы
Тема 6. Оценка моделей, ансамбли и введение в Deep Learning.	4				8			16	28	Отчет по лабораторной работы
ИТОГО за семестр:	18				36			90	144	Зачет с оценкой

Таблица 3 - Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых в них компетенций

Разделы, темы дисциплины (модуля)	Кол-во часов	Код компетенции		общее количество компетенций
		ПК-3	ПК-5	
Тема 1. Введение в Data Science и методология KDD (Knowledge Discovery in Databases)	20	+	+	2
Тема 2. Подготовка и предварительный анализ данных (Data Preparation & EDA).	20	+	+	2

Тема 3. Инженерия признаков (Feature Engineering).	20	+	+	2
Тема 4. Обучение с учителем (Supervised Learning).	28	+	+	2
Тема 5. Обучение без учителя (Unsupervised Learning).	28	+	+	2
Тема 6. Оценка моделей, ансамбли и введение в Deep Learning.	28	+	+	2
Итого	144	+	+	2

Краткое содержание каждой темы дисциплины (модуля)

Тема №1. Введение в Data Science и методология KDD

Определение понятий "Интеллектуальный анализ данных" (Data Mining), "Машинное обучение" (Machine Learning) и "Наука о данных" (Data Science). Связь с Big Data и Искусственным Интеллектом. Детальный разбор типовых задач машинного обучения: классификация (бинарная, многоклассовая), регрессия (прогнозирование численных значений), кластеризация (сегментация) и поиск ассоциативных правил. Изучение стандартных методологий ведения проектов: KDD (Knowledge Discovery in Databases) и CRISP-DM (Cross-Industry Standard Process for Data Mining), их основные этапы (понимание бизнеса, понимание данных, подготовка данных, моделирование, оценка, внедрение). Обзор ключевых инструментов и библиотек экосистемы Python: NumPy для вычислений, Pandas для манипуляции данными (DataFrame), Scikit-learn как основная библиотека ML, Matplotlib/Seaborn для визуализации.

Тема №2. Подготовка и предварительный анализ данных (Data Preparation & EDA)

Важность этапа подготовки данных ("Garbage In, Garbage Out"). Методы работы с "грязными" данными: стратегии обработки пропущенных значений (удаление, заполнение средним/медианой/модой, MICE). Техники обнаружения и обработки аномалий (выбросов), например, с использованием Z-оценки или межквартильного размаха (IQR). Исследовательский анализ данных (EDA): использование описательной статистики (среднее, медиана, дисперсия) для понимания распределения данных. Визуализация как основной инструмент EDA: гистограммы, диаграммы рассеяния (scatter plots), "ящички с усами" (box plots) для выявления связей и закономерностей. Преобразование признаков: методы кодирования категориальных данных (Label Encoding, One-Hot Encoding, Target Encoding). Необходимость масштабирования численных признаков (стандартизация и нормализация) для алгоритмов, чувствительных к масштабу (например, k-NN, SVM).

Тема №3. Инженерия признаков (Feature Engineering)

Определение инженерии признаков как процесса создания информативных переменных (признаков) из сырых данных для повышения производительности модели. Основные подходы: извлечение признаков (Feature Extraction), например, выделение дня недели/месяца из даты, или TF-IDF для текста. Создание признаков (Feature Creation) путем комбинации существующих (например, полиномиальные признаки). Отбор признаков (Feature Selection) для борьбы с "проклятием размерности", снижения шума и ускорения обучения. Разбор трех групп методов отбора: методы-фильтры (на основе корреляции, хи-квадрат), методы-обертки (Wrapper methods, например, рекурсивное исключение признаков) и встроенные методы (Embedded, например, L1-регуляризация в Lasso).

Тема №4. Обучение с учителем (Supervised Learning)

Фундаментальный принцип обучения с учителем: наличие размеченных данных (пары "объект-ответ"). Задачи регрессии: изучение Линейной регрессии (простой и множественной), понятие функции потерь (Loss Function), например, MSE (среднеквадратичная ошибка), и метод градиентного спуска (Gradient Descent) для оптимизации параметров. Задачи классификации: Логистическая регрессия для задач бинарной классификации (концепция логит-преобразования). Метод k-ближайших соседей (k-NN) как пример метрического (ленивого) алгоритма. Деревья решений (Decision Trees): критерии ветвления (Gini, Entropy). Метод опорных векторов (SVM): идея разделяющей гиперплоскости и "ядерного трюка" (kernel trick).

Тема №5. Обучение без учителя (Unsupervised Learning)

Обучение на неразмеченных данных с целью выявления скрытой структуры. Задачи кластеризации: K-Means (k-средних) как метод, основанный на центроидах, проблема выбора 'k' (метод "локтя"). Иерархическая (агломеративная) кластеризация и дендрограммы. DBSCAN как плотностной алгоритм, способный находить кластеры произвольной формы и выделять шум. Снижение размерности: Метод главных компонент (PCA) для уменьшения количества признаков с минимальной потерей информации, его отличие от отбора признаков. Поиск ассоциативных правил: анализ рыночной корзины, алгоритм Apriori, ключевые метрики (поддержка, достоверность, лифт).

Тема №6. Оценка моделей, ансамбли и введение в Deep Learning

Критическая важность оценки модели на отложенных данных (test set). Метрики качества: для классификации (Accuracy, Precision, Recall, F1-score, ROC-AUC) и регрессии (MAE, MSE, RMSE, R²). Проблемы обучения: Переобучение (Overfitting) и недообучение (Underfitting), методы диагностики (кривые обучения). Техника кросс-валидации (Cross-Validation) (K-Fold) для надежной оценки. Ансамблевые методы: Бэггинг (Bootstrap Aggregating) на примере Random Forest (Случайный лес). Бустинг (Boosting) на примере Gradient Boosting (GBM, XGBoost, LightGBM) как один из самых мощных методов для табличных данных. Введение в Deep Learning: базовые концепции нейронных сетей, перцептрон, функции активации, отличие глубокого обучения (автоматическая инженерия признаков) от классического ML.

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

Учебная деятельность студента в процессе изучения строится из контактных форм работы с преподавателем (лекции, лабораторные работы, зачет) и самостоятельной работы.

Лекция – это ведущая форма группового обучения. Ведущей она является потому, что именно с нее начинается изучение каждой новой дисциплины, темы. И только после лекции следуют другие, подчиненные ей формы обучения: семинары, практические занятия и т. д.

Методологическое значение лекции состоит в том, что в ней раскрываются фундаментальные теоретические основы учебной дисциплины и научные методы, с помощью которых анализируются жизненные явления. В целом можно сказать, что лекции как форме и методу обучения присущи три основные педагогические функции, которые определяют ее возможности и достоинства в учебном процессе: познавательная, развивающая и организующая.

Организационно-методической базой проведения лекционных занятий является рабочий учебный план направления или специальности. При подготовке лекционного материала преподаватель обязан руководствоваться учебными программами по дисциплинам кафедры,

тематика и содержание лекционных занятий которых представлена в рабочих программах. При чтении лекций преподаватель имеет право самостоятельно выбирать формы и методы изложения материала, которые будут способствовать качественному его усвоению. При этом в установленном порядке он может использовать технические средства обучения, имеющиеся на кафедре и в университете. Вместе с тем, всякий лекционный курс является в определенной мере авторским, представляет собой творческую переработку материала и неизбежно отражает личную точку зрения лектора на предмет и методы его преподавания. В этой связи представляется целесообразным привести некоторые общие методические рекомендации по построению лекционного курса и формам его преподавания.

Для успешного освоения дисциплины является обязательным посещение всех занятий, выполнение домашнего задания и иных форм самостоятельной работы, которые назначаются преподавателем.

Методическая поддержка дисциплины обеспечивается использованием дистанционных технологий. Студентам предлагается информационный ресурс, расположенный по адресу: <http://moodle.asu-edu.ru>, на сервере дистанционного обучения АГУ им. В. Н. Татищева. Доступ студентов к учебным ресурсам осуществляется по учетной записи и паролю после регистрации на курс «Интеллектуальный анализ данных и машинное обучение» на период обучения по данной дисциплине.

На сервере размещен методический материал по данной дисциплине, в содержание которого входит:

- теоретический материал;
- мультимедийные презентации по тематикам лекций;
- задания и указания по выполнению лабораторных работ, требования к содержанию и их оформлению, рекомендации по их защите;
- вопросы к экзамену.

Аудиторные занятия проводятся на основе теоретического материала, опубликованного на образовательном портале, это позволяет студентам изучить пропущенный материал или самостоятельно разобраться с темой, не освоенной на занятии.

5.2. Указания для обучающихся по освоению дисциплины (модулю)

Подготовка к опросу включает в себя:

- внимательное изучение материала лекций, относящихся к данному занятию, внимательное чтение учебного материала и дополнительной литературы;
- знакомство с терминологией;
- подготовка развернутых ответов на контрольные вопросы;
- определить трудные для понимания положения и подготовить по ним вопросы.

В рамках дисциплины «Экономика и Интеллектуальный анализ данных и машинное обучение» предполагается организация следующих видов самостоятельной работы студентов (таблица 4):

- работа с лекционным материалом, учебно-методическим информационным обеспечением;
- подготовка к лабораторным работам, подготовка к защите лабораторных работ;
- подготовка к зачету.

В качестве форм и методов контроля внеаудиторной самостоятельной работы используются: электронные отчеты по выполнению лабораторных работ; устный опрос, протоколы компьютерного тестирования.

Таблица 4 - Содержание самостоятельной работы обучающихся

Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Формы работы
Тема 1. Введение в Data Science и методология KDD (Knowledge Discovery in Databases) 1. Этические и правовые аспекты Data Science (конфиденциальность данных, предвзятость моделей (Bias)). 2. Сравнительный анализ методологий CRISP-DM, KDD и SEMMA.	14	Собеседование
Тема 2. Подготовка и предварительный анализ данных (Data Preparation & EDA). 1. Автоматизация EDA: библиотеки (например, Pandas Profiling, Sweetviz) и их применение. 2. Продвинутое методы заполнения пропусков (например, MICE, k-NN Imputer).	14	Собеседование
Тема 3. Инженерия признаков (Feature Engineering). 1. Особенности инженерии признаков для временных рядов (лаговые признаки, скользящие средние, сезонные компоненты). 2. Базовые методы векторизации текстов (Bag-of-Words, TF-IDF) как инженерия признаков.	14	Собеседование
Тема 4. Обучение с учителем (Supervised Learning). 1. Наивный Байесовский классификатор (Naive Bayes) и его применение. 2. Методы работы с несбалансированными выборками (Oversampling/Undersampling, SMOTE).	16	Собеседование
Тема 5. Обучение без учителя (Unsupervised Learning). 1. Алгоритмы обнаружения аномалий (Isolation Forest, Local Outlier Factor). 2. Методы нелинейного снижения размерности для визуализации (t-SNE, UMAP).	16	Собеседование
Тема 6. Оценка моделей, ансамбли и введение в Deep Learning. 1. Интерпретация моделей (Explainable AI): методы LIME и SHAP. 2. Сравнительный анализ библиотек градиентного бустинга (XGBoost, LightGBM, CatBoost) и их гиперпараметров.	16	Собеседование

5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины (модуля), выполняемые обучающимися самостоятельно

В процессе обучения студенты выполняют лабораторные работы. Результатом работы, выполняемой обучающимися, является электронный отчет по выполнению лабораторной работы.

Электронный отчет представляет собой файл формата doc, docx или pdf, содержащий программный код, результаты выполнения программы и текстовые пояснения. Файл передается на проверку преподавателю путем загрузки на ресурс <http://moodle.asu-edu.ru> в соответствующий заданию раздел.

Задания к лабораторным занятиям размещены на образовательном портале <http://moodle.asu-edu.ru>. Рекомендуется заранее ознакомиться с темой, основными вопросами и рекомендациями.

В процессе подготовки к лабораторным работам, необходимо обратить особое внимание на самостоятельное изучение рекомендованной литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала.

6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

6.1. Образовательные технологии

На лабораторных занятиях студент вначале знакомится с содержанием работы, пользуясь электронными методическими материалами, размещенными на <http://moodle.asu-edu.ru>, затем выполняет задание и показывает результаты преподавателю. Лабораторные работы выполняются студентом самостоятельно, возникающие при их выполнении проблемы разрешаются в рамках учебного времени и индивидуальных и групповых консультаций. Для выставления баллов по итогам выполнения ЛР студенты прикрепляют файлы с выполненными работами на образовательный портал.

В рамках организации самостоятельной работы студентам рекомендуется:

- работа с лекционным материалом;
- дополнительная подготовка к лабораторным работам или выполнение части лабораторной работы, которую они не успели сделать в аудитории;
- подготовка к текущей и промежуточной аттестации (экзамену).

Таблица 5. Образовательные технологии, используемые при реализации учебных занятий

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Тема 1. Введение в Data Science и методология KDD (Knowledge Discovery in Databases)	Обзорная лекция с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Тема 2. Подготовка и предварительный анализ данных (Data Preparation & EDA).	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Тема 3. Инженерия признаков (Feature Engineering).	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Тема 4. Обучение с учителем (Supervised Learning).	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Тема 5. Обучение без учителя (Unsupervised Learning).	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Тема 6. Оценка моделей, ансамбли и введение в Deep Learning.	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle

6.2. Информационные технологии

При реализации различных видов учебной и внеучебной работы используются следующие информационные технологии:

- использование возможностей интернета в учебном процессе (использование сайта преподавателя (рассылка заданий, предоставление выполненных работ, ответы на вопросы, ознакомление обучающихся с оценками и т.д.);
- использование электронных учебников и различных сайтов (например, электронных библиотек, журналов и т.д.) как источников информации;
- использование возможностей электронной почты преподавателя;
- использование средств представления учебной информации (электронных учебных пособий и практикумов, применение новых технологий для проведения очных (традиционных) лекций и семинаров с использованием презентаций и т.д.);
- использование интегрированных образовательных сред, где главной составляющей являются не только применяемые технологии, но и содержательная часть, т.е. информационные ресурсы (доступ к мировым информационным ресурсам, на базе которых строится учебный процесс);
- использование виртуальной обучающей среды (LMS Moodle «Электронное образование»)

6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
Adobe Reader	Программа для просмотра электронных документов
LMS Moodle	Образовательный портал ФГБОУ ВО «АГУ»
Microsoft Office	Пакет офисных программ
OpenOffice	Пакет офисных программ
7-zip	Архиватор
Microsoft Windows	Операционная система
Kaspersky Endpoint Security	Средство антивирусной защиты
Google Chrome	Браузер
Opera	Браузер
Anaconda Navigator	Графический интерфейс для работы с библиотеками Python

6.3.2. Современные профессиональные базы данных и информационные справочные системы

1. Электронная библиотека «Астраханский государственный университет» собственной генерации на платформе ЭБС «Электронный Читальный зал – БиблиоТех». <https://biblio.asu.edu.ru>.
2. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». <https://www.studentlibrary.ru>.
3. Электронный каталог Научной библиотеки АГУ на базе MARK SQL НПО «Информ-систем». <https://library.asu.edu.ru>.

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

7.1. Паспорт фонда оценочных средств

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) «Интеллектуальный анализ данных и машинное обучение» проверяется сформированность у

обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

Таблица 5 – Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

№ п/п	Контролируемые разделы дисциплины	Код контролируемой компетенции (компетенций)		Наименование оценочного средства
1.	Тема 1. Введение в Data Science и методология KDD (Knowledge Discovery in Databases)	ПК-3	ПК-5	Опрос
2.	Тема 2. Подготовка и предварительный анализ данных (Data Preparation & EDA).	ПК-3	ПК-5	Опрос
3.	Тема 3. Инженерия признаков (Feature Engineering).	ПК-3	ПК-5	Опрос
4.	Тема 4. Обучение с учителем (Supervised Learning).	ПК-3	ПК-5	Опрос
5.	Тема 5. Обучение без учителя (Unsupervised Learning).	ПК-3	ПК-5	Опрос
6.	Тема 6. Оценка моделей, ансамбли и введение в Deep Learning.	ПК-3	ПК-5	Опрос

Для оценивания результатов обучения в виде **знаний** используются следующие типы контроля:

- индивидуальное собеседование (устный опрос).
- электронный отчет.

Для оценивания результатов обучения в виде **умений и владений** используются практические контрольные задания (далее – ПКЗ), включающих одну или несколько задач (вопросов) в виде краткой формулировки действий (комплекса действий), которые следует выполнить, или описание результата, который нужно получить.

7.2. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

Критерии оценивания, используемые при компьютерном тестировании.

Оценка результатов компьютерного тестирования выполняется автоматически. Процент выполнения теста рассчитывается в зависимости от количества верных ответов по формуле:

$$\mathbf{0-59:2;60-79:3;80-89:4;90-100:5.}$$

Полученный процент выполнения переводится в балльную шкалу, в зависимости от установленного значения максимального балла за выполняемый тест (таблица 6). Перерасчет баллов осуществляется автоматически.

Критерии оценивания, используемые при отчете ЛР.

В системе Moodle балл за выполнение лабораторной работы выставляется в 100- балльной шкале комплексно с учетом степени подготовки студента к выполнению работы, объема выполненной работы на занятии и оформлении отчета в соответствии с перечисленными критериями. В зависимости от выставленного максимального балла (таблица 6) перерасчет за каждый отчет ЛР начисляемых баллов производится автоматически.

Таблица 6- Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
90-100	<ul style="list-style-type: none"> - содержание отчета соответствуют номеру варианта, выданного преподавателем - задания выполнены правильно - задания выполнены в полном объеме - информация изложена достоверно, обоснованно, логично, последовательно - информация представлена иллюстративно - продемонстрировано отличное владение инструментальными средствами обработки информации - отчет оформлен в соответствии с требованиями ГОСТ - отчет представлен в установленные сроки
80-89	<ul style="list-style-type: none"> - содержание отчета соответствуют номеру варианта, выданного преподавателем - задания выполнены правильно, но присутствуют некоторые неточности - задания выполнены в полном объеме - информация изложена достоверно, но есть нарушения в последовательности и логичности ее изложения - информация представлена иллюстративно - продемонстрировано хорошее владение инструментальными средствами обработки информации - отчет оформлен в соответствии с требованиями ГОСТ - отчет представлен в установленные сроки
60-79	<ul style="list-style-type: none"> - содержание отчета соответствуют номеру варианта, выданного преподавателем - задания выполнены правильно, но присутствуют ошибки - задания выполнены в объеме не менее 60% - информация изложена достоверно, но есть нарушения в последовательности и логичности ее изложения - информация представлена не иллюстративно - продемонстрировано удовлетворительное владение инструментальными средствами обработки информации - отчет оформлен в соответствии с требованиями ГОСТ, но с некоторыми незначительными нарушениями - отчет представлен в установленные сроки
0-59	<ul style="list-style-type: none"> - содержание отчета соответствуют номеру варианта, выданного преподавателем - задания выполнены с ошибками - задания выполнены в объеме менее 60% - информация изложена не достоверно, в последовательности и логичности изложения допущены существенные ошибки - информация представлена не иллюстративно - продемонстрировано неудовлетворительное владение инструментальными средствами обработки информации - отчет оформлен в соответствии с требованиями ГОСТ, имеются существенные нарушения - отчет не представлен, или представлен с нарушением срока сдачи без уважительной причины

Таблица 7 - Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
9-10	Студент свободно владеет теоретическим материалом, может характеризовать теоретические аспекты на основе практических примеров, ответ отличается профессиональной культурой, даны полные и верные ответы на дополнительные вопросы.
7-8	Студент владеет теоретическим материалом, ответ логичен, изложение теоретического материала сопровождается практическими примерами, имеются отдельные негрубые ошибки, при ответе на дополнительные вопросы допущены небольшие неточности.
5-6	Студент владеет теоретическим материалом, но в изложении отсутствует логика, имеются существенные недочеты, отсутствуют практические примеры к излагаемым теоретическим вопросам, при ответе на дополнительные вопросы допущены неточности.
0-4	Студент не владеет теоретическим материалом или неверно определяет основные профессиональные понятия, не даны ответы на дополнительные вопросы.

Грубыми считаются ошибки, свидетельствующие о том, что студент:

- не овладел основным материалом дисциплины
- не может применять на практике полученные знания
- не знает формул, графиков, схем и т.п.
- не знает приемов решения задач, аналогичных ранее решенным.

Не грубыми ошибками являются

- неточность чертежа, графика, схемы и т.п.
- неточно сформулированный вопрос или пояснение при решении задачи
- отдельные ошибки вычислительного характера

Недочетами считаются

- отдельные погрешности в формулировке вопроса или ответа
- отдельные ошибки вычислительного характера
- небрежное выполнение записей, чертежей, схем, графиков и т.п.

7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)

Лабораторная работа №1. Знакомство с инструментами Data Science и процессом KDD

- **Цель:** Освоение базовой рабочей среды (Jupyter Notebook / Google Colab) и основных библиотек (Pandas, NumPy) для загрузки и первичного осмотра данных.
- **Задачи:**
 1. Настроить рабочую среду (локально или в облаке).
 2. Изучить основы работы с Pandas DataFrame.
 3. Загрузить набор данных (например, из CSV-файла).
 4. Провести первичный "осмотр" данных: оценить размер, типы столбцов, наличие пропусков.
- Краткое содержание задания:

Студенту предоставляется набор данных (например, "Титаник" или "Цены на недвижимость"). Необходимо загрузить данные в Jupyter Notebook, используя библиотеку Pandas. Вывести на экран первые 10 строк, последние 10 строк, общую информацию о данных (.info()) и основные описательные статистики для числовых столбцов (.describe()).

Лабораторная работа №2. Исследовательский анализ данных (EDA) и предобработка

- **Цель:** Научиться выявлять и обрабатывать проблемы в данных (пропуски, выбросы) и строить визуализации (с помощью Matplotlib/Seaborn) для формирования гипотез.
- **Задачи:**
 1. Проанализировать пропущенные значения и выбрать стратегию их заполнения (удаление, заполнение медианой/модой).
 2. Построить базовые графики (гистограммы, диаграммы рассеяния, "ящики с усами") для понимания распределения данных.
 3. Построить матрицу корреляций (heatmap) для выявления связей между признаками.
 4. Выполнить кодирование категориальных признаков (Label Encoding / One-Hot Encoding).
- **Краткое содержание задания:**

На основе набора данных из Лаб. 1: 1) Провести анализ пропущенных значений и применить адекватную стратегию их заполнения (обосновав выбор). 2) Построить гистограммы для 2-3 числовых признаков и bar plot для 1-2 категориальных. 3) Построить матрицу корреляций Пирсона. 4) Преобразовать категориальные признаки в численный вид, готовый для моделирования.

Лабораторная работа №3. Инженерия признаков (Feature Engineering)

- **Цель:** Освоить техники создания новых (синтетических) признаков и методы отбора признаков для улучшения качества будущей модели.
- **Задачи:**
 1. Научиться создавать новые признаки путем комбинации или трансформации существующих.
 2. Применить масштабирование данных (StandardScaler или MinMaxScaler).
 3. Изучить и применить простой метод отбора признаков (например, на основе корреляции или Feature Importance из дерева решений).
 4. Разделить итоговый набор данных на обучающую и тестовую выборки (train/test split).
- **Краткое содержание задания:**

Используя обработанный на Лаб. 2 набор данных: 1) Создать как минимум 2 новых признака (например, 'РазмерСемьи' = 'Родители' + 'Дети', или извлечь 'Титул' из 'Имени'). 2) Выполнить масштабирование всех числовых признаков. 3) Разделить данные на обучающую (train) и тестовую (test) выборки в соотношении 80/20.

Лабораторная работа №4. Построение моделей обучения с учителем

- **Цель:** Научиться применять, обучать и сравнивать базовые алгоритмы обучения с учителем (например, Логистическая регрессия и Дерево решений) на подготовленных данных.
- **Задачи:**
 1. Определить тип задачи (классификация или регрессия).
 2. Обучить 2-3 простые модели (например, LogisticRegression, DecisionTreeClassifier, k-NN) на обучающей выборке.
 3. Получить предсказания моделей на тестовой выборке.
 4. Рассчитать базовые метрики качества (например, Accuracy) и сравнить модели.
- **Краткое содержание задания:**

На данных, подготовленных в Лаб. 3: 1) Обучить модель Логистической регрессии. 2) Обучить модель Дерева решений (с ограничением глубины, например max_depth=5). 3) Для каждой модели получить предсказания на тестовой выборке. 4) Рассчитать метрику 'Accuracy' (для классификации) или 'MSE' (для регрессии) и сделать вывод, какая модель показала себя лучше на данной задаче.

Лабораторная работа №5. Применение методов обучения без учителя (Кластеризация)

- **Цель:** Освоить применение алгоритма K-Means для сегментации данных и научиться интерпретировать полученные кластеры.
- **Задачи:**
 1. Подготовить данные для кластеризации (выбрать признаки, масштабировать).

2. Использовать "метод локтя" (Elbow method) для подбора оптимального количества кластеров (k).
 3. Обучить модель K-Means с найденным 'k'.
 4. Проанализировать характеристики полученных кластеров (например, средние значения признаков в каждом кластере).
- Краткое содержание задания:
Используя набор данных (можно взять новый, например, "Mall Customers", или использовать только признаки из "Титаника", игнорируя целевую переменную): 1) Выбрать 2-3 числовых признака для кластеризации. 2) Построить график "метода локтя" для k от 1 до 10. 3) Выбрать оптимальное 'k' и запустить K-Means. 4) Визуализировать кластеры (если 2 признака) и проанализировать центроиды: описать, чем в среднем отличаются объекты в разных кластерах.

Лабораторная работа №6. Оценка качества и ансамблевые методы

- **Цель:** Углубленно изучить метрики качества и применить ансамблевые методы (Random Forest) для повышения точности прогнозирования.
- **Задачи:**
 1. Рассчитать расширенные метрики качества для моделей из Лаб. 4 (Матрица ошибок, Precision, Recall, F1-score).
 2. Обучить ансамблевую модель (например, RandomForestClassifier).
 3. Сравнить метрики ансамбля с метриками простых моделей.
 4. Применить K-Fold кросс-валидацию (Cross-Validation) для более надежной оценки качества лучшей модели.
- Краткое содержание задания:
Вернуться к задаче из Лаб. 4: 1) Для лучшей модели из Лаб. 4 рассчитать матрицу ошибок, Precision, Recall и F1-score. 2) Обучить модель Случайного леса (Random Forest) на тех же данных. 3) Сравнить все метрики Случайного леса с метриками Логистической регрессии/Дерева решений. 4) Провести K-Fold кросс-валидацию (на 5 фолдах) для Случайного леса, чтобы оценить стабильность модели, и рассчитать среднюю ассурасу.

Полный пакет лабораторных работ размещен на сайте moodle.asu-edu.ru на курсе дисциплины.

Перечень вопросов и заданий, выносимых на зачёт

1. Перечислите и кратко опишите этапы методологии KDD (Knowledge Discovery in Databases).
2. В чем заключается разница между задачами классификации, регрессии и кластеризации? Приведите по одному примеру для каждой.
3. Какова цель Исследовательского анализа данных (EDA)? Назовите 3-4 основных инструмента (графика), используемых в EDA.
4. Опишите как минимум три стратегии обработки пропущенных значений (пропусков) в наборе данных.
5. Зачем необходимо кодирование категориальных признаков? Сравните методы One-Hot Encoding и Label Encoding, указав их плюсы и минусы.
6. В каких случаях необходимо применять масштабирование (Standardization) и нормализацию (Normalization) данных? Какие алгоритмы к этому чувствительны?
7. Что такое «инженерия признаков» (Feature Engineering) и почему этот этап считается одним из важнейших в ML?
8. Опишите три основные группы методов отбора признаков: методы-фильтры (Filter), методы-обертки (Wrapper) и встроенные (Embedded).
9. В чем заключается «проклятие размерности» (Curse of Dimensionality)?
10. Опишите суть метода главных компонент (PCA). Является ли он методом отбора признаков?

11. Объясните принцип работы линейной регрессии. Какая функция потерь (Loss Function) в ней обычно используется?
12. В чем разница между Линейной и Логистической регрессией? Для какого типа задач используется каждая из них?
13. Как работает алгоритм k-ближайших соседей (k-NN)? Какие у него есть преимущества и недостатки?
14. Опишите процесс построения Древа решений (Decision Tree). Что такое критерии ветвления (Gini, Entropy)?
15. Объясните базовый принцип работы метода опорных векторов (SVM). Что такое "разделяющая гиперплоскость"?
16. Опишите по шагам алгоритм кластеризации K-Means.
17. Как подобрать оптимальное количество кластеров (k) для алгоритма K-Means? Опишите «метод локтя» (Elbow Method).
18. В чем основное отличие иерархической кластеризации от K-Means?
19. В чем преимущество алгоритма DBSCAN по сравнению с K-Means?
20. Для чего используется поиск ассоциативных правил (алгоритм Apriori)? Объясните метрики: Поддержка (Support) и Достоверность (Confidence).
21. Что такое переобучение (Overfitting) и недообучение (Underfitting) модели? Назовите способы борьбы с переобучением.
22. Что показывает матрица ошибок (Confusion Matrix)?
23. Опишите метрики Precision (Точность), Recall (Полнота) и F1-score. В каких случаях использование Accuracy (Доля правильных ответов) может ввести в заблуждение?
24. Зачем нужна кросс-валидация (Cross-Validation)? Опишите принцип работы K-Fold CV.
25. Что такое ансамблевые методы? В чем фундаментальная разница между бэггингом (Bagging) и бустингом (Boosting)?
26. Как работает алгоритм Случайного леса (Random Forest)?
27. Опишите основную идею Градиентного бустинга (Gradient Boosting).
28. Что такое "гиперпараметры" модели? Приведите примеры.
29. Какие методы используются для интерпретации "черных ящиков" (Explainable AI)? (Например, LIME или SHAP).
30. Назовите ключевые компоненты искусственной нейронной сети (нейрон, слой, функция активации).

Таблица 9 – Примеры оценочных средств с ключами правильных ответов

№п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-3. Способен выполнять разработку программного обеспечения общего и прикладного назначения, проводить оценку качества и работоспособности разработанного программного обеспечения, осуществлять его сопровождение и модификацию				
1.	Задания закрытого типа	Задача определения, является ли email спамом или нет, — это пример: А) Задачи регрессии В) Задачи кластеризации С) Задачи классификации Д) Задачи снижения размерности	С	1 мин
2.		Какой метод используется для преобразования категориального признака (например, "Цвет": Красный, Синий, Зеленый) в несколько бинарных (0/1) столбцов? А) Стандартизация	В	1 мин

		(Standardization) B) One-Hot Encoding C) Label Encoding D) K-Means		
3.		Какой из этих алгоритмов является методом обучения БЕЗ учителя (Unsupervised Learning)? A) Логистическая регрессия B) Дерево решений C) K-Means (Метод k-средних) D) Случайный лес (Random Forest)	C	1 мин
4.	Задания открытого типа	Опишите этапы методологии CRISP-DM	Методология CRISP-DM (Cross-Industry Standard Process for Data Mining) включает 6 фаз: 1. Business Understanding (Понимание бизнеса): Определение целей и требований проекта с точки зрения бизнеса. 2. Data Understanding (Понимание данных): Сбор, первичный анализ и исследование данных. 3. Data Preparation (Подготовка данных): Очистка, преобразование и форматирование данных (включая инженерию признаков) для моделирования. 4. Modeling (Моделирование): Выбор и применение техник машинного обучения, настройка гиперпараметров. 5. Evaluation (Оценка): Оценка качества модели(ей) с точки зрения бизнес-целей; проверка, достигнуты ли цели. 6. Deployment (Внедрение): Интеграция модели в бизнес-процессы или системы.	5 мин
5.		Сравните методы кодирования категориальных признаков: One-Hot Encoding и Label Encoding. В каких случаях предпочтительнее использовать каждый из них?	Label Encoding (Кодирование метками): Преобразует категории в числа (например, "Низкий"=0, "Средний"=1, "Высокий"=2). - Предпочтительно: Для порядковых (ordinal) данных, где важен порядок (как в примере). Также его можно использовать с моделями, основанными на	5 мин

			<p>деревьях (например, Random Forest), которые не воспримут эти числа как линейную зависимость.</p> <ul style="list-style-type: none"> - One-Hot Encoding (Прямое кодирование): Создает N новых бинарных (0/1) столбцов, где N — количество уникальных категорий. - _Предпочтительно:_ Для номинальных данных, где нет внутреннего порядка (например, "Красный", "Синий", "Зеленый"). Этот метод обязателен для линейных моделей (Линейная/Логистическая регрессия, SVM), так как Label Encoding создал бы ложную числовую связь между категориями. 	
6.		<p>В чем заключается фундаментальное различие между задачами регрессии и задачами классификации в обучении с учителем?</p>	<p>Основное различие — в типе целевой переменной (того, что мы прогнозируем):</p> <ul style="list-style-type: none"> - В классификации целевая переменная дискретная (категориальная). Модель предсказывает класс или метку (например, "спам" / "не спам", "кошка" / "собака", "дефолт" / "не дефолт"). - В регрессии целевая переменная непрерывная (числовая). Модель предсказывает числовое значение (например, цену дома, температуру воздуха, спрос на товар). 	5 мин
7.	<p>Задание комбинированного типа</p>	<p>Вам предоставлен набор данных о клиентах телеком-компании (длительность контракта, ежемесячный платеж, объем трафика, возраст) и целевая переменная, указывающая, ушел клиент (1) или остался (0).</p> <p>Часть 1 (Закрытая): К какому типу задач машинного обучения относится этот сценарий?</p> <p>A) Регрессия B) Кластеризация C) Классификация D) Поиск ассоциативных правил</p> <p>Часть 2 (Открытая): Назовите два различных алгоритма, которые</p>	<p>Часть 1 (Закрытая): C) Классификация (Задача состоит в том, чтобы предсказать дискретную метку класса — "уйдет" или "останется").</p> <p>Часть 2 (Открытая): 1. Логистическая регрессия: Подходит, потому что это классический, быстрый и хорошо интерпретируемый алгоритм, разработанный специально для задач бинарной</p>	10 мин

		можно использовать для решения этой задачи, и кратко объясните, почему они подходят.	классификации_ (когда есть ровно два класса: 0 или 1). 2. Случайный лес (Random Forest): Подходит, так как это мощный ансамблевый метод (на основе деревьев решений), который хорошо справляется с табличными данными, устойчив к выбросам, не требует тщательного масштабирования признаков и способен выявлять сложные нелинейные зависимости между признаками и целевой переменной.	
ПК-5. Способен вести сдачу проекта, собирать и анализировать мнения и замечания заказчика по выполнению проекта и предлагать соответствующие решения				
1.	Задания закрытого типа	Ситуация, когда модель показывает высокое качество на обучающих данных, но низкое качество на новых (тестовых) данных, называется: А) Недообучение (Underfitting) В) Переобучение (Overfitting) С) Кросс-валидация (Cross-Validation) D) Нормализация (Normalization)	В	1 мин
2.		Какая метрика измеряет долю объектов, которые были корректно классифицированы как "положительные" (True Positives), среди всех объектов, которые фактически являются положительными? А) Точность (Precision) В) Полнота (Recall) С) Доля верных ответов (Accuracy) D) F1-мера (F1-Score)	В	1 мин
3.		К какому типу ансамблевых методов относится алгоритм «Случайный лес» (Random Forest), который обучает множество деревьев на случайных подвыборках данных? А) Бустинг (Boosting) В) Бэггинг (Bagging) С) Стекинг (Stacking) D) Метод главных компонент (PCA)	В	1 мин
4.	Задания открытого типа	Что такое переобучение (Overfitting) модели? Назовите 2-3 способа борьбы с ним.	Переобучение — это ситуация, когда модель машинного обучения становится излишне сложной и "запоминает" обучающие данные, включая шум и случайные	5 мин

			<p>выбросы, вместо того, чтобы выучить общую закономерность. В результате модель показывает очень высокое качество на обучающей выборке, но очень низкое качество на новых (тестовых) данных.</p> <p>Способы борьбы:</p> <ol style="list-style-type: none"> 1. Упрощение модели: (например, уменьшение глубины дерева решений, уменьшение числа нейронов/слоев в нейросети). 2. Регуляризация (L1/L2): Добавление "штрафа" к функции потерь за большие веса признаков (для линейных моделей, нейросетей). 3. Кросс-валидация: Для более надежной оценки качества и подбора гиперпараметров. 4. Сбор большего количества данных: Помогает модели лучше отделить сигнал от шума. 5. Feature Selection: Отбор только наиболее важных признаков, удаление шума. 	
5.		<p>Объясните назначение метрик "Точность" (Precision) и "Полнота" (Recall) и приведите пример, когда одна из них важнее другой</p>	<p>- Precision (Точность): Показывает, какая доля объектов, которые модель назвала "положительными", действительно являются таковыми. (Фокус: минимизация False Positives).</p> <p>- Recall (Полнота): Показывает, какую долю реально "положительных" объектов модель смогла обнаружить. (Фокус: минимизация False Negatives).</p> <p>- Пример, где важен Precision: Спам-фильтры. Мы не хотим, чтобы важное письмо (False Positive) попало в спам. Лучше пропустить немного спама (ниже</p>	5 мин

			<p>Recall), чем потерять важное письмо (выше Precision).</p> <p>- Пример, где важен Recall: Медицинская диагностика (например, рака). Мы хотим найти <u>всех</u> больных (высокий Recall). Мы готовы мириться с тем, что здоровым людям (False Positives) назначат доп. обследование, но мы не можем пропустить больного (False Negative).</p>	
6.		Сравните два основных ансамблевых подхода: Бэггинг (Bagging) и Бустинг (Boosting)	<p>Бэггинг (Bagging) (e.g., Random Forest):</p> <p>- Принцип: Модели (например, деревья) обучаются параллельно и независимо друг от друга.</p> <p>- Данные: Каждая модель обучается на своей случайной подвыборке данных (бутстрэп).</p> <p>- Результат: Результаты усредняются (для регрессии) или определяется голосованием (для классификации).</p> <p>- Цель: Снижение дисперсии (variance) (уменьшение переобучения).</p> <p>Бустинг (Boosting) (e.g., Gradient Boosting, XGBoost):</p> <p>- Принцип: Модели обучаются последовательно.</p> <p>- Данные: Каждая <u>следующая</u> модель фокусируется на ошибках, допущенных <u>предыдущей</u> моделью, пытаясь их исправить.</p> <p>- Результат: Модели взвешенно суммируются.</p> <p>- Цель: Снижение смещения (bias) (повышение общей точности).</p>	5мин
7.	Задание комбинированного типа	Модель для диагностики редкого заболевания была протестирована на 1000 пациентах. Из 50 <u>фактически</u> больных пациентов, модель правильно	<p>Часть 1 (Закрытая): C)</p> <p>81.8% (45 / 55)</p> <p>- Расчет: Precision = TP / (TP + FP) = 45 / (45 + 10) = 45 / 55 ≈ 81.8%.</p>	10 мин

	<p>определила 45 (TP=45, FN=5). При этом 10 здоровым пациентам (из 950) модель ошибочно поставила диагноз "болен" (FP=10, TN=940). Часть 1 (Закрытая): Рассчитайте Precision (Точность) модели. (Precision = TP / (TP + FP)) А) 90.0% (45 / 50) В) 98.9% (45 / 45.5) С) 81.8% (45 / 55) D) 97.8% (45 / 46) Часть 2 (Открытая): В этом медицинском сценарии, какая метрика является более критичной — Precision или Recall — и почему?</p>	<p>- Пояснение к другим вариантам: 90.0% (45/50) — это Recall (TP / (TP+FN)). Часть 2 (Открытая): Более критичной является Recall (Полнота). - Почему: В медицинской диагностике (особенно редких, но опасных заболеваний) самое важное — не пропустить больного (минимизировать False Negatives). Высокий Recall (в данном случае 45/50 = 90%) означает, что мы нашли почти всех, кто болен. Низкий Precision (81.8%) означает, что мы получили несколько ложноположительных результатов (FP=10), то есть здоровых людей отправили на дообследование. Это нежелательно, но гораздо менее опасно, чем пропустить реальный случай заболевания (FN=5).</p>	
--	--	--	--

Полный комплект оценочных материалов по дисциплине (модулю) (фонд оценочных средств) хранится в электронном виде на кафедре, утверждающей рабочую программу дисциплины (модуля).

7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)

Таблица 10. Технологическая карта рейтинговых баллов по дисциплине (модулю)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	Ответ на занятии	4/5	20	
2.	Выполнение практического задания (лабораторная работа.)	6/10	60	
3.	Тестирование	1/0,5 балла за каждый правильный ответ	10	
Всего			90	-
Блок бонусов				
4.	Отсутствие пропусков занятий	2,5	2,5	
5.	Своевременное выполнение всех заданий	3	3	
6.	Активное участие в семинарском занятии, обсуждении семинарских вопросов	9/0,5	4,5	
Всего			10	-

ИТОГО	100	-
--------------	------------	----------

Таблица 11 – Система штрафов (для одного занятия)

Показатель	Балл
<i>Опоздание на занятие</i>	-1
<i>Нарушение учебной дисциплины</i>	-1
<i>Неготовность к занятию</i>	-3
<i>Пропуск занятия без уважительной причины</i>	-2

Таблица 12 – Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)

Сумма баллов	Оценка по 4-балльной шкале	
90–100	5 (отлично)	Зачтено
85–89	4 (хорошо)	
75–84		
70–74		
65–69	3 (удовлетворительно)	
60–64		
Ниже 60	2 (неудовлетворительно)	Не зачтено

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

а) Основная литература:

1. Жерон, О. Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow : [концепции, инструменты и методы для создания интеллектуальных систем] / Орельен Жерон ; [пер. с англ. А. А. Слинкина]. – 3-е изд. – Москва : ДМК Пресс, 2023. – 976 с. – ISBN 978-5-97060-896-1. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-896-1/>.
2. Воронцов, К. В. Математические методы обучения по прецедентам : курс лекций / К. В. Воронцов. – Москва : МФТИ, 2011. – 118 с. – URL: <http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-Logic-slides.pdf>.
3. Вандерплас, Д. Python для сложных задач : наука о данных и машинное обучение / Д. Вандерплас ; [пер. с англ. А. В. Логунова]. – Санкт-Петербург : Питер, 2023. – 576 с. – ISBN 978-5-4461-1502-7. – URL: <https://www.litres.ru/book/d-vanderplas/python-dlya-slozhnyh-zadach-nauka-o-dannyh-i-mashinnoe-obuchenie-68931252/>.
4. Мюллер, А. Введение в машинное обучение с помощью Python : руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо ; [пер. с англ. А. Б. Снастина]. – 2-е изд. – Москва : ДМК Пресс, 2022. – 482 с. – ISBN 978-5-97060-848-0. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-848-0/>.
5. Рассел, С. Искусственный интеллект : современный подход / С. Рассел, П. Норвиг ; [пер. с англ.]. – 4-е изд. – Москва : Диалектика, 2022. – 1088 с. – ISBN 978-5-907458-00-8. – URL: <https://www.ozon.ru/product/iskusstvennyy-intellekt-sovremennyy-podhod-rassel-styuart-norvig-piter-473133800/>.
6. Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. – Москва : Интернет-Университет Информационных Технологий : БИНОМ. Лаборатория знаний, 2008. – 382 с. – ISBN 978-5-94774-817-8. – URL: <https://www.ozon.ru/product/data-mining-uchebnoe-posobie-14819448/>.

7. Хан, Д. Data Mining : концепции и методы / Д. Хан, М. Камбер, Д. Пэй ; [пер. с англ.]. – 3-е изд. – Москва : Диалектика, 2020. – 704 с. – ISBN 978-5-907144-80-9. – URL: <https://www.labyrinth.ru/books/744081/> .

б) Дополнительная литература:

7. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, Й. Бенджио, А. Курвилль ; [пер. с англ. А. А. Слинкина]. – 2-е изд. – Москва : ДМК Пресс, 2021. – 652 с. – ISBN 978-5-97060-851-0. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-851-0/> .
8. Саттон, Р. С. Обучение с подкреплением / Р. С. Саттон, Э. Дж. Барто ; [пер. с англ. А. А. Слинкина]. – 2-е изд. – Москва : ДМК Пресс, 2020. – 552 с. – ISBN 978-5-97060-843-5. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-843-5/> .
9. Бишоп, К. М. Распознавание образов и машинное обучение / К. М. Бишоп ; [пер. с англ. А. С. Крючковой]. – Москва : Диалектика, 2023. – 944 с. – ISBN 978-5-907695-15-1. – URL: <https://www.ozon.ru/product/raspoznavanie-obrazov-i-mashinnoe-obuchenie-911855661/> .
10. Мерфи, К. П. Машинное обучение : вероятностный подход / К. П. Мерфи ; [пер. с англ. В. М. Кулюкина]. – Москва : ДМК Пресс, 2024. – 1018 с. – ISBN 978-5-97060-826-8. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-826-8/> .
11. Паклин, Н. Б. Data Science : наука о данных : [учебный курс] / Н. Б. Паклин, В. И. Орешков. – Санкт-Петербург : Питер, 2022. – 512 с. – ISBN 978-5-4461-1221-7. – URL: <https://www.litres.ru/book/v-i-oreshkov/data-science-nauka-o-dannyh-67455823/> .
12. Хасти, Т. Элементы статистического обучения : интеллектуальный анализ данных, логический вывод и прогнозирование / Т. Хасти, Р. Тибширани, Д. Фридман ; [пер. с англ.]. – 2-е изд. – Москва : ДМК Пресс, 2022. – 792 с. – ISBN 978-5-97060-825-1. – URL: <https://dmkpress.com/catalog/electronics/data-analysis-and-data-mining/978-5-97060-825-1/> .
13. Дьяконов, А. Г. Введение в машинное обучение в Python / А. Г. Дьяконов. – [Б. м.] : ДМК Пресс, 2020. – 302 с. – ISBN 978-5-97060-822-0. – URL: <https://www.ozon.ru/product/vvedenie-v-mashinnoe-obuchenie-v-python-181519448/> .
14. Аббот, С. Аналитика для руководителей : интеллектуальный анализ данных для принятия бизнес-решений / С. Аббот ; [пер. с англ. Ю. С. Лаврова]. – Санкт-Петербург : Питер, 2022. – 368 с. – ISBN 978-5-4461-1632-1. – URL: <https://www.litres.ru/book/sin-abbot/analitika-dlya-rukovoditeley-intellektualnyy-analiz-dannyh-d-66952737/> .

8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)

1. Электронная библиотека «Астраханский государственный университет» собственной генерации на платформе ЭБС «Электронный Читальный зал – БиблиоТех».
2. <https://biblio.asu.edu.ru> Учетная запись образовательного портала АГУ.
3. Многопрофильный образовательный ресурс «Консультант студента» является электронной библиотечной системой, предоставляющей доступ через сеть Интернет к учебной литературе и дополнительным материалам, приобретенным на основании прямых договоров с правообладателями. Каталог в настоящее время содержит около 15000 наименований.
4. www.studentlibrary.ru. Регистрация с компьютеров АГУ
5. Электронная библиотечная система издательства ЮРАЙТ, раздел «Легендарные книги».
6. www.biblio-online.ru

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Для проведения занятий по дисциплине имеются лекционные аудитории, оборудованные мультимедийной техникой с возможностью презентации обучающих материалов, фрагментов фильмов; аудитории для проведения семинарских и практических занятий, оборудованные

учебной мебелью и средствами наглядного представления учебных материалов; библиотека с местами, оборудованными компьютерами, имеющими доступ к сети Интернет.

10. ОСОБЕННОСТИ РЕАЛИЗАЦИИ ДИСЦИПЛИНЫ (МОДУЛЯ) ПРИ ОБУЧЕНИИ ИНВАЛИДОВ И ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. При обучении лиц с ограниченными возможностями здоровья учитываются их индивидуальные психофизические особенности. Обучение инвалидов осуществляется также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).

Для лиц с нарушением слуха возможно предоставление учебной информации в визуальной форме (краткий конспект лекций; тексты заданий, напечатанные увеличенным шрифтом), на аудиторных занятиях допускается присутствие ассистента, а также сурдопереводчиков и тифлосурдопереводчиков. Текущий контроль успеваемости осуществляется в письменной форме: обучающийся письменно отвечает на вопросы, письменно выполняет практические задания. Доклад (реферат) также может быть представлен в письменной форме, при этом требования к содержанию остаются теми же, а требования к качеству изложения материала (понятность, качество речи, взаимодействие с аудиторией и т. д.) заменяются на соответствующие требования, предъявляемые к письменным работам (качество оформления текста и списка литературы, грамотность, наличие иллюстрационных материалов и т. д.). Промежуточная аттестация для лиц с нарушениями слуха проводится в письменной форме, при этом используются общие критерии оценивания. При необходимости время подготовки к ответу может быть увеличено.

Для лиц с нарушением зрения допускается аудиальное предоставление информации, а также использование на аудиторных занятиях звукозаписывающих устройств (диктофонов и т. д.). Допускается присутствие на занятиях ассистента (помощника), оказывающего обучающимся необходимую техническую помощь. Текущий контроль успеваемости осуществляется в устной форме. При проведении промежуточной аттестации для лиц с нарушением зрения тестирование может быть заменено на устное собеседование по вопросам. Для лиц с ограниченными возможностями здоровья, имеющих нарушения опорно-двигательного аппарата, на аудиторных занятиях, а также при проведении процедур текущего контроля успеваемости и промежуточной аттестации могут быть предоставлены необходимые технические средства (персональный компьютер, ноутбук или другой гаджет); допускается присутствие ассистента (ассистентов), оказывающего обучающимся необходимую техническую помощь (занять рабочее место, передвигаться по аудитории, прочитать задание, оформить ответ, общаться с преподавателем).