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	УТВЕРЖДАЮ Зав. каф. информационных технологий
А. В. Григорьев	А. Н. Марьенков
«04» апреля 2024 г.	«04» апреля 2024 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«БОЛЬШИЕ ДАННЫЕ»

Составитель(и)	**Кузнецова В. Ю., доцент, к.т.н., доцент кафедры ИТ;**
Согласовано с работодателями:	**Машкова Е. Ю., заместитель руководителя Управления Федеральной Службы Государственной Статистики по Астраханской области и республике Калмыкия;** **Кособрюхова Т. Н., руководитель службы записи актов гражданского состояния Астраханской области;**
Направление подготовки / специальность	**39.03.01 СОЦИОЛОГИЯ**
Направленность (профиль) / специализация ОПОП	**АНАЛИТИКА БОЛЬШИХ ДАННЫХ В СОЦИАЛЬНЫХ ПРОЦЕССАХ**
Квалификация (степень)	**бакалавр**
Форма обучения	**очная**
Год приёма	**2024**
Курс	**2**
Семестр(ы)	**4**

Астрахань – 2024

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целью освоения дисциплины (модуля) «Большие данные» является изучение математических методов и моделей, используемых в системах обработки и анализа больших данных для поддержки принятия решений, и развитие профессиональных навыков в этой области.

1.2. Задачи освоения дисциплины (модуля):

- сформировать представление о проблемах анализа и обработки данных;
- сформироватьнавыкиразработкиалгоритмованализаиобработкиданныхсприменением моделей DataMining.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) «Большие данные» относится к части, формируемой участниками образовательных отношений, и осваивается в 3 семестре.

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

– Введение в информационные технологии

Знания: базовые понятия информатики и вычислительной техники; понятие информационной системы и информационной технологии; технические и программные средства реализации информационных процессов; основные устройства, входящие в состав ЭВМ, их назначение и характеристики; формы представления и преобразования информации в компьютере.

Умения: применять вычислительную технику для решения практических задач; разработать алгоритм поставленной задачи.

Навыки: работы на персональном компьютере.

– Введение в программирование

Знания: основные структуры данных, используемые в языках программирования; структуру программ; основные принципы алгоритмизации.

Умения: создавать схему алгоритма для задачи; проводить отладку и тестирование созданного программного продукта.

Навыки в области алгоритмизации, разработки, отладки и тестирования программных продуктов.

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

- Большие данные в социологии.
- Алгоритм обработки больших данных.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс освоения дисциплины (модуля) направлен на формирование элементов следующей(их) компетенции(ий) в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки / специальности:

ПК-5: Способен собирать, обрабатывать и анализировать большие данные в исследованиях социальных процессов

Таблица 1. Декомпозиция результатов обучения

Код компетенции	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине (модулю)		
		Знать (1)	Уметь (2)	Владеть (3)
ПК-5	ПК-5: Способен собирать, обрабатывать и анализировать большие данные в исследованиях социальных процессов	Знает способы сбора, обработки и анализа больших данных	Умеет собирать, обрабатывать и анализировать большие данные	Владеет навыками использования больших данных в исследованиях социальных процессов
		Знает специфику использования результатов анализа больших данных в социологии	Умеет применять базовые теоретические знания в области больших данных в научных и научно-прикладных исследованиях социальных процессов.	Владеет алгоритмами обработки больших данных

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Общая трудоемкость дисциплины в соответствии с учебным планом составляет 2 зачетные единицы (72 часа).

Трудоемкость отдельных видов учебной работы студентов очной, очно-заочной и заочной форм 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	2
Объем дисциплины в академических часах	72
Контактная работа обучающихся с преподавателем (всего), в том числе (час.):	36
- занятия лекционного типа, в том числе:	18
- практическая подготовка (если предусмотрена)	1
- занятия семинарского типа (семинары, практические, лабораторные), в том числе:	18
- практическая подготовка (если предусмотрена)	1
Самостоятельная работа обучающихся (час.)	36
Форма промежуточной аттестации обучающегося	зачет – 4 семестр

Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий и самостоятельной работы, для каждой 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.							СР, час.	Итого часов	Форма текущего контроля успеваемости, форма промежуточной аттестации
	Л		ПЗ		ЛР		КР / КП			
	Л	в т.ч. ПП	ПЗ	в т.ч. ПП	ЛР	в т.ч. ПП				
Тема 1. Большие данные и экосистема больших данных	2	0	0	0	4	0	0	4	10	Лабораторная работа, Устный опрос
Тема 2. Процесс исследования данных	2	0	0	0	2	0	0	5	9	Лабораторная работа, Устный опрос
Тема 3. Машинное обучение и математические основы работы с данными	2	0	0	0	2	0	0	4	8	Лабораторная работа, Устный опрос
Тема 4. Описательная статистика	2	0	0	0	2	0	0	5	9	Лабораторная работа, Устный опрос
Тема 5. Теория вероятностей при работе с большими данными	2	0	0	0	2	0	0	4	8	Лабораторная работа, Устный опрос
Тема 6. Проверка гипотез при анализе данных	2	0	0	0	2	0	0	5	9	Лабораторная работа, Устный опрос
Тема 7. Визуализация больших данных	2	0	0	0	2	0	0	4	8	Лабораторная работа, Устный опрос
Итоговый проект	4	1	0	0	2	1	0	5	13	Проект на основе лабораторных работ
ИТОГО за семестр:	**18**	**1**	**0**	**0**	**18**	**1**	**0**	**36**	**72**	**ЗАЧЁТ**

Таблица 3. Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых компетенций

Раздел, тема дисциплины (модуля)	Кол-во часов	Код компетенции	Общее количество компетенций
		ПК-5	
Тема 1. Большие данные и экосистема больших данных	15	+	1
Тема 2. Процесс исследования данных	13	+	1
Тема 3. Машинное обучение и математические основы работы с данными	13	+	1
Тема 4. Описательная статистика	13	+	1
Тема 5. Теория вероятностей при работе с большими данными	13	+	1
Тема 6. Проверка гипотез при анализе данных	13	+	1

Раздел, тема дисциплины (модуля)	Кол-во часов	Код компетенции	Общее количество компетенций
		ПК-5	
Тема 7. Визуализация больших данных	13	+	1
Итоговый проект	15	+	1
Итого	108	+	1

Далее приводится краткое содержание каждой темы дисциплины (модуля)]

Тема 1. Большие данные и экосистема больших данных.

Ведение в анализ больших данных. Основные определения, термины, задачи анализа больших данных. Экосистема аналитики больших данных. Распределенные файловые системы.

Тема 2. Процесс исследования данных

Методы и методики процесса исследования данных, управление проектами в сфере аналитики данных.

Тема 3. Машинное обучение и математические основы работы с данными

Машинное обучение на больших данных. Обзор источников информации для Big Data. Методики сбора данных.

Тема 4. Описательная статистика

Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных. Описательная статистика имеющихся наборов данных и инструменты для этого.

Тема 5. Теория вероятностей при работе с большими данными

Событие. Случайная величина. Центральная предельная теорема. Распределение вероятностей. Закон больших чисел. Применение при работе с большими данными.

Тема 6. Проверка гипотез при анализе данных

Статистические тесты. Дисперсионный анализ. Критерий Хи-квадрат. Технические детали.

Тема 7. Визуализация больших данных

Создание дашбордов. Построение информационных панелей с использованием современных инструментов

Итоговый проект

Анализ технического задания на проведение исследование. Интервьюирование заказчика. Выполнение самостоятельного рабочего проекта. Составление отчёта по проекту

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 (аудиторные занятия, экзамен) и самостоятельной работы.

Для успешного освоения дисциплины является обязательным посещение всех занятий, выполнение самостоятельной работы, которая 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. На сервере размещен методический материал по данной дисциплине, в содержание которого входит:

– теоретический материал;

– задания и указания по выполнению лабораторных работ.

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)

В рамках дисциплины «Большие данные» предполагается организация следующих видов самостоятельной работы студентов:

– работа с учебно-методическим информационным обеспечением;

– подготовка к лабораторным работам, подготовка отчетов.

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

Таблица 4. Содержание самостоятельной работы обучающихся

Вопросы, выносимые на самостоятельное изучение	Кол-во часов	Форма работы
Тема 1. Большие данные и экосистема больших данных. Элементы экосистемы больших данных.	10	Отчет о выполнении ЛР 1 Устный опрос
Тема 2. Процесс исследования данных. Apache Hadoop и его элементы. Межотраслевой стандартный процесс для исследования данных (CRISP-DM).Элементы процесса исследования данных.	9	Отчет о выполнении ЛР 2 Устный опрос
Тема 3. Машинное обучение и математические основы работы с данными. Типы задач машинного обучения. Примеры алгоритмов.	8	Отчет о выполнении ЛР 3 Устный опрос
Тема 4. Работа с большими наборами данных. MapReduce. Как работает MapReduce. Пример кода (псевдокода).	9	Отчет о выполнении ЛР 4 Устный опрос
Тема 5. Hadoop и Spark. Подготовка данных в Spark. Работа со Spark в интерактивном режиме. Сохранение данных в Hive. Построение интерактивного отчета и дашборда с помощью платформы Qlik Sense.	8	Отчет о выполнении ЛР 5 Устный опрос
Тема 6. Базы данных NoSQL. Использование Neo4j. Выполнение запросов на Cypher. Создание узла пользователя в базе данных Neo4j. Визуализация найденных закономерностей, построение графов.	9	Отчет о выполнении ЛР 6 Устный опрос
Тема 7. Интеллектуальный анализ текста как источника больших данных. Алгоритм выделения основы. Классификация с использованием наивного баесовского классификатора и дерева принятия решений. Тренировка и оценка моделей. Визуализация выявленных закономерностей.	8	Отчет о выполнении ЛР 7 Устный опрос
Тема 8. Визуализация больших данных. Построение графиков и диаграмм с использованием библиотеки dc.js. Создание информационной панели в браузере. Работа с библиотекой визуализации данных d3.js.	13	Отчет о выполнении ЛР 8 Устный опрос

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

В рамках реализации компетентностного подхода в соответствии с требованиями ФГОС ВО в учебном процессе предусмотрены активные и интерактивные формы проведения занятий. Основой для выстраивания аудиторных занятий служит технология развития критического мышления, которая, интегрируя элементы проблемного, проектного, дискуссионного обучения, позволяет достигать максимальной эффективности в достижении проектируемых компетенций.

Цели дисциплины достигаются путем сочетания контактной и самостоятельной работы студентов: лабораторных занятий на ПК и организации самостоятельной работы студентов.

Лабораторные работы выполняются студентами с применением ПК и ориентированы на формирование деятельностных компетентностей. Они заключаются в выполнении сквозного цикла лабораторных работ. В процессе выполнения лабораторных работ достигаются следующие цели:

- изучается экосистема работы с большими данными;
- формируются практические навыки работы с большими массивами информации при решении конкретных практических задач;
- формируется навык выявления ошибочных и нештатных ситуаций и реагирования на них.

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 и отчеты на образовательный портал.

Для самостоятельного изучения теоретического материала дисциплины рекомендуется использовать интернет-ресурсы, информационные базы, методические разработки, специальную учебную и научную литературу.

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
Большие данные и экосистема больших данных	Обзорная лекция с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Процесс исследования данных	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Машинное обучение и математические основы работы с данными	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Описательная статистика	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Теория вероятностей при работе с большими данными	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Проверка гипотез при анализе данных	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Визуализация больших данных	Лекция-диалог с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Итоговый проект	Лекция-диалог с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle

6.2. Информационные технологии

При реализации учебной и внеучебной работы используются следующие информационные технологии:

– образовательный портал http://moodle.asu.edu.ru (размещение учебно-методического материала, публикация заданий для предоставления студентами выполненных отчетов по всем видам работ, ознакомление учащихся с оценками и т.д., размещение объявлений, обсуждение вопросов в форуме и т.д.), как элемента интерактивного взаимодействия участников образовательного процесса (технологии дистанционного обучения);

– среда разработки моделей машинного обучения https://colab.research.google.com

– веб-сервис для хостинга IT-проектов и их совместной разработки https://github.com

– ресурсы ЭБС и сети Internet, как источников информации.

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) «Большие данные» 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 6. Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

Контролируемый раздел, тема дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Большие данные и экосистема больших данных	ПК-5	Лабораторная работа, Устный опрос
Процесс исследования данных	ПК-5	Лабораторная работа, Устный опрос
Машинное обучение и математические основы работы с данными	ПК-5	Лабораторная работа, Устный опрос
Описательная статистика	ПК-5	Лабораторная работа, Устный опрос

Контролируемый раздел, тема дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Теория вероятностей при работе с большими данными	ПК-5	Лабораторная работа, Устный опрос
Проверка гипотез при анализе данных	ПК-5	Лабораторная работа, Устный опрос
Визуализация больших данных	ПК-5	Лабораторная работа, Устный опрос
Итоговый проект	ПК-5	Проект на основе лабораторных работ

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Таблица 7. Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

Таблица 8. Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание по подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задания

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

Перечень вопросов и заданий, выносимых на зачёт

1. Что такое большие данные?
2. Основные характеристики больших данных («три V»).
3. Отличия Big Data от традиционных баз данных.
4. Классификация источников больших данных.
5. Этапы жизненного цикла анализа больших данных.
6. Какие технологии входят в экосистему Hadoop?
7. Чем отличаются HDFS и MapReduce?
8. Что такое Apache Spark и его основные компоненты?
9. Основные этапы процесса исследования данных (CRISP-DM).
10. Что включает этап понимания бизнеса в CRISP-DM?
11. Описание этапа подготовки данных.
12. Почему важно проведение разведочного анализа данных?
13. Примеры методов очистки данных.
14. Зачем необходима нормализация данных?
15. Основные типы моделей машинного обучения.
16. Принцип работы алгоритма линейной регрессии.
17. Алгоритм k ближайших соседей — область применения и ограничения.
18. Основы метода главных компонент (PCA).
19. Что такое кластерный анализ и для чего он используется?
20. Как работает алгоритм случайного леса?
21. Основное назначение градиентного бустинга.
22. Особенности глубокого обучения и его применение.
23. Метрики оценки классификации (accuracy, precision, recall, F1 score).
24. Регуляризация в моделях машинного обучения.
25. Перекрёстная проверка (cross validation): цель и способы реализации.
26. Центральные тенденции выборки (среднее арифметическое, медиана, мода).
27. Меры разброса данных (дисперсия, стандартное отклонение, квартили).
28. Что показывает коэффициент вариации?
29. Виды диаграмм распределения данных (гистограмма, box plot).
30. Для чего используют ковариацию и корреляцию?
31. Как визуализируются многомерные данные?
32. Интерквартильный размах и выбросы данных.
33. Преимущества графического представления статистики перед численными показателями.
34. Вероятностные события и пространство элементарных исходов.
35. Условная вероятность и независимость событий.
36. Закон больших чисел и центральная предельная теорема.
37. Нормальное распределение и его свойства.
38. Распределение Пуассона и экспоненциальное распределение.
39. Байесовская теория вероятности и её применение.
40. Метод Монте-Карло и его использование в моделировании.
41. Роль случайных процессов в анализе временных рядов.
42. Выбор подходящего распределения для реальных данных.
43. Применение вероятностных подходов в рекомендательных системах.
44. Статистическая значимость результатов тестирования гипотезы.
45. Типичные ошибки первого и второго рода при проверке гипотез.

46. Критерии принятия решений при нулевых и альтернативных гипотезах.
47. Т-тест и ANOVA тест: различия и области применения.
48. Тест хи-квадрат: принцип и условия применимости.
49. Методы множественного сравнения и контроль ложноположительных выводов.
50. Практика интерпретации p-value.
51. Использование доверительного интервала при оценке среднего значения.
52. A/B тестирование и принципы анализа экспериментальных данных.
53. Ограничения статистической значимости при анализе больших объемов данных.
54. Цели и задачи визуализации больших данных.
55. Какой тип графика лучше всего отображает временную динамику?
56. Как эффективно представить категориальные переменные?
57. Основные подходы к интерактивной визуализации данных.
58. Тепловая карта и её использование в анализе зависимостей признаков.
59. Чего позволяет достичь визуализация высокомерных данных?
60. Проблемы восприятия сложных графиков и рекомендации по улучшению наглядности.
61. Трендовые линии и линии скользящего среднего: особенности применения.
62. Дашбординг и создание отчетов на основе больших данных.

Тематика и краткое содержание лабораторных работ

Полная версия лабораторных работ представлена на платформе Moodle.

Тема 1. Большие данные и экосистема больших данных
Лабораторная работа 1

Большие данные и Интернет Вещей.

Работа с платформой IBM Cloud. Построение аналитических потоков.

Тема 2. Процесс исследования данных
Лабораторная работа 2

Работа с большими наборами данных.

Задача: прогнозирование вредоносных веб-адресов. Использование разреженного представления данных. Использование сжатых данных вместо необработанных. Применение онлайнового алгоритма для прогнозирования. Построение рекомендательной системы внутри базы данных.

Метод k-ближайших соседей. Метод локально-чувствительного хеширования. Метрики расстояния.

Тема 3. Машинное обучение и математические основы работы с данными
Лабораторная работа 3

Использование Hadoop для аналитики больших данных.

Настройка Vagrant. Работа с подходом Map Reduce.

Тема 4. Описательная статистика
Лабораторная работа 4

Использование Hadoop и Spark.

Задача: оценка риска при кредитовании. Работа в среде vagrant на виртуальной машине. Взаимодействие с HDFS. Загрузка и сохранение данных в Hadoop. Подготовка данных в Spark. Работа со Spark в интерактивном режиме. Сохранение данных в Hive.

Тема 5. Теория вероятностей при работе с большими данными
Лабораторная работа 5

Настройка виртуальной машины. Распространение переменных по всем узлам кластера. Широковещательные переменные. Аккумуляторные переменные. Предобработка данных в среде Spark. Работа с пропущенными данными. RDD. Работа с объектами DataFrame.

Тема 6. Проверка гипотез при анализе данных

Лабораторная работа 6

Построение поисковой системы диагностики болезней на базе Elasticsearch.

Сбор данных и индексирование болезней. Диагностика болезней по симптомам. Обработка ошибок и опечаток с помощью метрики расстояния Дамерау-Левенштейна. Задача профилирования болезни через агрегирование значимых терминов. Отображение информации.

Построение рекомендательной системы для связанных данных с использованием графовой базы данных. Задача: построение рекомендаций рецептов блюд на основе предпочтений пользователей и набора ингредиентов. Использование индекса Elasticsearch для чистки данных и для заполнения графовой базы данных. Использование Neo4j. Выполнение запросов на Cypher. Создание узла пользователя в базе данных Neo4j. Визуализация найденных закономерностей, построение графов.

Классификация с использованием наивного байесовского классификатора и дерева принятия решений. Тренировка и оценка моделей. Визуализация выявленных закономерностей.

Тема 7. Визуализация больших данных

Лабораторная работа 7

Создание информационной панели (дашборда).

Использование Crossfilter для фильтрации набора данных с применением MapReduce. Построение графиков и диаграмм с использованием библиотеки dc.js. Создание информационной панели в браузере. Работа с библиотекой визуализации данных d3.js.

Итоговый проект

Выполнение итогового проекта по реальным данным от заказчика. Анализ технического задания. Составление отчёта.

Таблица 9. Примеры оценочных средств с ключами правильных ответов

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-5				
1.	Задание закрытого типа	Документ-ориентированной СУБД являются: 1. Oracle Database 2. MS SQL Server 3. Mongo DB 4. SQLite	3	3
2.		Графовой СУБД является 1. MS SQL Server 2. Mongo DB 3. Neo4j 4. SQLite	3	3
3.		Какое значение пропущено в выражении «Матрица с разреженностью больше … является разреженной матрицей» 1. 0,4	2	3

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
		2. 0,5 3. 0,6 4. 0,9		
4.		Какой инструмент визуализации НЕ является облачным? 1. SAP Analytics Cloud, 2. Google Data Studio, 3. Yandex Datalens, 4. Tableau	4	3
5.		Какое свойство НЕ обязательно для больших данных? 1. Ценность данных 2. Структурированность данных 3. Объем 4. Скорость накопления	2	3
6.	Задание открытого типа	Какая матрица называется разреженной?	Разреженная матрица - это матрица, в которой большинство элементов равно нулю. Напротив, таблица, в которой большинство элементов отличны от нуля, называется плотной. Мы определяем разреженность матрицы как число нулевых элементов, деленное на общее количество элементов. Матрица с разреженностью больше 0,5 является разреженной матрицей.	5
7.		Дайте определение усиковой диаграмме.	Усиковая диаграмма - график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.	5
8.		Что может быть отображено на усиковой диаграмме?	Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.	5
9.		Дайте определение пузырьковой диаграмме.	Пузырьковая диаграмма — это разновидность	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			точечной диаграммы, в которой точки данных заменены пузырьками, причем их размер служит дополнительным измерением данных. На пузырьковой диаграмме, как и на точечной, нет оси категорий — и горизонтальная, и вертикальная оси являются осями значений. В дополнение к значениям X и значениям Y, наносимым на точечную диаграмму, на пузырьковой диаграмме показаны также значения Z (размер).	
10.		В чем отличие пузырьковой диаграммы от точечной.	Вы можете использовать пузырьковую диаграмму вместо точечной, если данные состоят из трех рядов, каждый из которых содержит набор значений. Размеры пузырьков определяются значениями третьего ряда данных. Пузырьковые диаграммы часто используются для представления финансовых данных. Пузырьки разных размеров позволяют визуально выделить конкретные значения.	5

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)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	*Ответ на занятии*	16/1	16	По расписанию
2.	*Выполнение практической работы*	6/9	54	
3.	*Выполнение контрольной работы*	2/5	10	

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
4.	*Тест*	1/5	5	
5.	*Реферат*	1/5	5	
Всего			**90**	-
Блок бонусов				
6.	*Посещение занятий без пропусков*	1	3	
7.	*Своевременное выполнение всех заданий*	1	3	
8.	*Активность студента на занятии*	1	4	
Всего			**10**	-
ИТОГО			**100**	-

Таблица 11 – Система штрафов (для одного занятия)

Показатель	Балл
Опоздание на занятие	- 1
Нарушение учебной дисциплины	- 1
Неготовность к занятию	- 2
Пропуск занятия без уважительной причины	- 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)	
85–89	4 (хорошо)	
75–84		
70–74		
65–69	3 (удовлетворительно)	
60–64		
Ниже 60	2 (неудовлетворительно)	

Зачёт проходит в форме устного собеседования со студентом по билетам, составленным из вопросов (п. 7.3). Один билет включает в себя 2 вопроса. Выбор билета осуществляется в случайном порядке. На подготовку студенту отводится не менее 40 мин. Во время проведения экзамена студенту запрещено пользоваться сотовым телефоном и иными средствами связи, персональным компьютером, сетью Интернет, заготовленными заранее ответами и т.п

Для стимулирования развития творческого и научно-исследовательского потенциала студентов при промежуточном оценивании предусмотрена система дополнительных баллов, а именно начисление до 10 поощрительных баллов за участие в конференциях, семинарах, выставках и т.п. в области анализа данных, программировании с представлением индивидуальных проектов в области аналитики данных.

Начисление баллов зависит от статуса мероприятия и статуса участия в нем студента. Начисление баллов происходит при предоставлении диплома, сертификата, грамоты, материалов конференции, опубликованной статьи, тезисов и т.п.

Преподаватель, реализующий дисциплину, в зависимости от уровня подготовленности обучающихся может использовать иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

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

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)

8.1. Основная литература

1. Броневич, А. Г. Нечеткие модели анализа данных и принятия решений : учебное пособие / А. Г. Броневич, А. Е. Лепский. - Москва : Высшая школа экономики, 2022. - 266 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". - ISBN 978-5-7598-2407-7. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785759824077.html

2. Искусственный интеллект, аналитика и новые технологии / - Москва : Альпина Паблишер, 2022. - 200 с. (Серия "Harvard Business Review: 10 лучших статей") - ISBN 978-5-9614-4791-0. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785961447910.html

3. Кошкаров А.В. Аналитика больших данных. Астрахань: Издатель Сорокин Роман Васильевич, 2018. URL: https://biblio.asu.edu.ru/Reader/Book/2019100910013323100002066826. (Электронная библиотека "Астраханский государственный университет")

4. Маккинли, У. Python и анализ данных / Уэс Маккинли - Москва : ДМК Пресс, 2015. - 482 с. - ISBN 978-5-97060-315-4. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970603154.html

5. Ульман, Дж. Д. Анализ больших наборов данных / Дж. Д. Ульман, Ю. Лесковец, А. Раджараман; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". - ISBN 978-5-89818-304-2. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785898183042.html

8.2. Дополнительная литература

1. Адлер Ю.П. Статистическое управление процессами. "Большие данные". М.: МИСиС, 2016. - 52 с. URL: http://www.studentlibrary.ru/book/ISBN9785876239693.html. (ЭБС «Консультант студента»).

2. Форман Дж. Много цифр: Анализ больших данных при помощи Excel. М.: Альпина Паблишер, 2016. - 461 с. URL : http://www.studentlibrary.ru/book/ISBN9785961450323.html. (ЭБС «Консультант студента»).

3. Будылдина Н.В. Сетевые технологии высокоскоростной передачи данных : Учебное пособие для вузов. М. : Горячая линия - Телеком, 2016. - 342 с. URL : http://www.studentlibrary.ru/book/ISBN9785991205368.html. (ЭБС «Консультант студента»).

4. Бэнкер К. MongoDB в действии. М.: ДМК Пресс, 2012. - 394 с. URL : http://www.studentlibrary.ru/book/ISBN9785940748311.html (ЭБС «Консультант студента»).

5. Кук Д. Машинное обучение с использованием библиотеки H2O. М. : ДМК Пресс, 2018. - 250 с. URL : http://www.studentlibrary.ru/book/ISBN9785970605080.html. (ЭБС «Консультант студента»).

6. Рашка С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения. М. : ДМК Пресс, 2017. - 418 с. URL : http://www.studentlibrary.ru/book/ISBN9785970604090.html. (ЭБС «Консультант студента»).

7. Сухов К.К. Node.js. Путеводитель по технологии. М.: ДМК Пресс, 2015. - 416 с. URL : http://www.studentlibrary.ru/book/ISBN9785970601648.html. (ЭБС «Консультант студента»).

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

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». www.studentlibrary.ru.

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Для проведения лабораторных занятий необходима аудитория, оснащенная компьютерными рабочими местами студентов и доступом в Интернет.

Для проведения **лекционных занятий**:

1. Используется аудитория, оборудованная необходимым количеством столов, стульев, доской маркерной и электронной.
2. Аудитория должна иметь следующие нормы освещенности

- СНиП 23-05-95 «Естественное и искусственное освещение» норма освещенности аудиторий ВУЗов 400 Лк.
- СанПиН 2.2.1/2.1.1.1278-03 «Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий» пункт 3.3.3. «Общее освещение в помещениях общественных зданий должно быть равномерным».

3. Электронная доска должна быть подключена к сети Интернет.

Для проведения **лабораторных занятий**:

1. Лабораторные занятия проводятся с группами или подгруппами не более 15 человек.
2. Аудитория должна быть оснащена необходимым количеством столов, стульев, доской маркерной и электронной.
4. Аудитория должна иметь следующие нормы освещенности

- СНиП 23-05-95 «Естественное и искусственное освещение» норма освещенности аудиторий ВУЗов 400 Лк.
- СанПиН 2.2.1/2.1.1.1278-03 «Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий» пункт 3.3.3. «Общее освещение в помещениях общественных зданий должно быть равномерным».

5. В аудитории должно быть не менее 15 компьютеров, находящихся в исправном состоянии.
6. Расположение компьютеров в аудитории должно позволять преподавателю подойти к рабочему месту студента.
7. Компьютеры должны быть соединены локальной сетью со скоростью не менее 1 Гбит/с и подключены к сети Интернет.
8. Компьютеры должны обладать минимальными характеристиками:

- Материнская плата H610M H DDR 4
- Процессор 12th Gen Intel(R) Core(TM) i3-12100
- Видеоадаптер Intel(R) UHD Graphics 730

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).