

УТВЕРЖДАЮ
Декан факультета

_____ Матвеев П.В.

« ____ » _____ 20__

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Направление/специальность подготовки	09.03.04 Программная инженерия
Специализация/профиль/программа подготовки	Разработка программно-информационных систем
Уровень высшего образования	Бакалавриат
Форма обучения	Очная
Факультет	О Естественнонаучный
Выпускающая кафедра	О7 Информационные системы и программная инженерия
Кафедра-разработчик рабочей программы	О7 Информационные системы и программная инженерия

КУРС	СЕМЕСТР	ОБЩАЯ ТРУДОЁМКОСТЬ (ЗАЧЕТНЫХ ЕДИНИЦ)	ЧАСЫ (по наличию видов занятий)									ВИД ПРОМЕЖУТОЧНОГО КОНТРОЛЯ
			ОБЩАЯ ТРУДОЁМКОСТЬ	АУДИТОРНЫЕ ЗАНЯТИЯ				САМОСТОЯТЕЛЬНАЯ РАБОТА				
				ВСЕГО	ЛЕКЦИИ	ЛАБОРАТОРНЫЙ ПРАКТИКУМ	ПРАКТИЧЕСКИЕ ЗАНЯТИЯ	ВСЕГО	КУРСОВОЙ ПРОЕКТ	КУРСОВАЯ РАБОТА	ДРУГИЕ ВИДЫ САМОСТ. РАБОТЫ	
3	5	4	144	68	34	0	34	76	0	0	76	диф. зач.

ЛИСТ СОГЛАСОВАНИЯ

**РАБОЧАЯ ПРОГРАММА СОСТАВЛЕНА В СООТВЕТСТВИИ С ТРЕБОВАНИЯМИ ФЕДЕРАЛЬНОГО
ГОСУДАРСТВЕННОГО ОБРАЗОВАТЕЛЬНОГО СТАНДАРТА ВЫСШЕГО ОБРАЗОВАНИЯ (ФГОС ВО)**

09.03.04 Программная инженерия

год набора группы: 2025

Программу составил:

Кафедра О7 Информационные системы и программная инженерия
Ярошевская Елена Юрьевна, старший преподаватель

Программа рассмотрена
на заседании кафедры-разработчика
рабочей программы **О7 Информационные системы и программная инженерия**

Заведующий кафедрой Семенова Е.Г., д.т.н., проф.

Программа рассмотрена
на заседании выпускающей кафедры

О7 Информационные системы и программная инженерия

Заведующий кафедрой Семенова Е.Г., д.т.н., проф.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Разделы рабочей программы

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ
2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО
3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ
4. ФОРМЫ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ
5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ
6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Приложения к рабочей программе дисциплины

- Приложение 1. Аннотация рабочей программы
- Приложение 2. Технологии и формы обучения
- Приложение 3. Фонды оценочных средств

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью освоения дисциплины является формирование следующих компетенций:

ПК-1.4 — Способен использовать технологии ИИ на различных этапах жизненного цикла программных продуктов, проектировать интеллектуальные компоненты программного обеспечения

ПК-93 — Способен генерировать новые идеи для решения задач цифровой экономики, абстрагироваться от стандартных моделей, перестраивать сложившиеся способы решения задач, выдвигать альтернативные варианты действий с целью выработки новых оптимальных алгоритмов

Формированию компетенций служит достижение следующих результатов образования:

ПК-1.4

знания:

основные законы распределения случайных величин;

виды вариационных рядов;

классификация гипотез и методы их проверки;

показатели описательной статистики и их интерпретация применительно к исследуемому набору данных;

показатели динамики;

умения:

организовать сбор репрезентативных данных в процессе исследования;

выбирать и применять адекватные методы анализа данных, полученных в результате исследований;

делать выводы по результатам применения статистических методов анализа;

навыки:

грамотно использовать прикладные программные пакеты для решения задач анализа больших данных;

визуализировать полученные результаты применения статистических методов анализа.

ПК-93

знания:

концепции проблемы и архитектурные подходы к обработке больших данных;

алгоритмы обработки больших данных при помощи технологий Spark, Kafka, Spark Streaming;

особенности адаптации алгоритмов на графах и алгоритмов машинного обучения к возможности обработки больших объемов данных;

умения:

применять алгоритмы машинного обучения и алгоритмы на графах для анализа больших данных;

писать SQL и NoSQL запросы для обработки данных на распределенном вычислительном кластере;

навыки:

проектирования и реализации пайплайнов обработки больших данных;

визуализировать результаты обработки больших данных.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

Дисциплина **МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ** является дисциплиной **части, формируемой участниками образовательных отношений блока 1**, программы подготовки по направлению **09.03.04 Программная инженерия**.

Содержание дисциплины является логическим продолжением дисциплин: **ОСНОВЫ СИСТЕМНОГО АНАЛИЗА, ТЕОРИЯ ВЕРОЯТНОСТЕЙ И НАЧАЛА МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ, ВВЕДЕНИЕ В ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ**.

Содержание дисциплины является основой для освоения дисциплин: **БАЗЫ ДАННЫХ, НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА (ПОЛУЧЕНИЕ ПЕРВИЧНЫХ НАВЫКОВ НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЫ), ПОСТРОЕНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ**.

Предварительные компетенции, сформированные у обучающегося до начала изучения дисциплины:

- ОПК-1 — Способен применять естественнонаучные и общетехнические знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности
- ОПК-2 — Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности
- ОПК-3 — Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности
- ПК-94 — Способен к управлению информацией и данными, поиску источников информации и данных, восприятию, анализу, запоминанию и передаче информации с использованием цифровых средств, а также с помощью алгоритмов при работе с полученными из различных источников данными с целью эффективного использования полученной информации для решения задач
- УК-1 — Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины составляет 4 з.е., 144 ч.

3.1. Содержание (дидактика) дисциплины

КУРС	СЕМЕСТР	Наименование разделов и дидактических единиц	ВСЕГО	Аудиторные занятия в контактной форме			Самостоятельная работа студентов	Формируемая компетенция, %	
				ВСЕГО	Лекции	Практические занятия		ПК-1.4	ПК-93
3	5	Раздел 1. Введение в теорию больших данных. Характеристики Big Data. Источники больших данных. Сбор больших данных. Архитектурные парадигмы. Российская экосистема.	12	6	4	2	6	10	20
3	5	Раздел 2. Хранение и и управление большими данными. Распределенные файловые системы. NoSQL базы данных. Инструменты хранения данных в Big Lakes.	16	6	4	2	10	20	20
3	5	Раздел 3. Пакетная обработка больших данных. Статистические методы в распределенной среде. Выборочные методы. Описательные статистики. Методы оценки связи между признаками. Методы сравнения выборок.	42	18	8	10	24	20	10
3	5	Раздел 4. Потокковая обработка больших данных. Основы потоковой обработки. Apache Spark Straming. ПО для исследования и визуализации Apache Superset. Динамические модели.	18	12	6	6	6	20	20
3	5	Раздел 5. Машинное обучение и статистический анализ больших данных. Проверка гипотез. Оценка качества моделей. Методы контроля качества. Методы снижения размерности. Анализ динамики.	38	18	8	10	20	20	10
3	5	Раздел 6. Оптимизация производительности. Оркестрация и управление пайплайнами. Мониторинг и отладка приложений. Профилирование задач. Интеграция со специализированными российскими ML-платформами.	18	8	4	4	10	10	20
Всего за 5 семестр			144	68	34	34	76	100	100
Всего по дисциплине			144	68	34	34	76	100	100

3.2. Аудиторный практикум

№ п/п	Номер и наименование раздела дисциплины	Тема практического занятия	Объем, ауд. часов
1	Раздел 1. Введение в теорию больших данных.	Сбор и первичная обработка данных с Росстата. Сырые данные и метаданные. Сравнительный анализ источников.	2
2	Раздел 2. Хранение и управление большими данными.	Создание распределенного хранилища. Организация Data Lake. Валидация целостности данных.	2
3	Раздел 3. Пакетная обработка больших данных.	Основные характеристики распределения и графическое представление. Ряды распределения. Дискретные и интервальные вариационные ряды. Гистограмма, полигон, кумулята.	2
4		Абсолютные и относительные показатели вариации. Понятие и способы расчета дисперсии, ее свойства. Правило сложения дисперсий. Способы обнаружения грубых погрешностей: критерий Романовского, критерий "трех сигм".	2
5		Задачи и условия применения корреляционного анализа. Параметрические и непараметрические методы оценки корреляции. Оценка силы связи между альтернативными признаками: коэффициенты ассоциации и контингенции. Линейная и нелинейные виды корреляционной зависимости.	2
6		Понятие парной корреляции, расчет и пределы изменения парного коэффициента корреляции, расчет и интерпретация парного коэффициента детерминации. Уравнение регрессии.	2
7		Уравнение множественной регрессии. Факторный анализ. Дисперсионный анализ. Кластерный анализ.	2
8	Раздел 4. Потокковая обработка больших данных.	Построение сквозного пайплайна. Настройка Spark Streaming. Анализ метрик.	3
9		Визуализация в Apache Superset. Построение дашбордов. Классы диаграмм.	3
10	Раздел 5. Машинное обучение и статистический анализ больших данных.	Классификация гипотез. Критерии принятия решения. Проверка гипотез о генеральной средней и равенстве двух выборочных средних. Проверка гипотез о виде распределения генеральной совокупности.	2
11		Расчет контрольных границ для построения контрольных карт Шухарта по индивидуальным значениям, средним, среднеквадратичным отклонениям и размаху.	2
12		Методы механического и аналитического сглаживания временных рядов, уравнение тренда; методы оценки качества уравнения тренда.	2

13	Раздел 6. Оптимизация производительности.	Моментные и интервальные временные ряды, элементы временного ряда. Абсолютные приросты, коэффициенты роста, темпы роста, темпы прироста, показатели средних.	2
14		Оценка связи между динамическими рядами, понятие ложной корреляции, методы исключения автокорреляции в рядах динамики.	2
15		Оркестрация ETL-пайплайнов с гарантией завершения. Интеграция с ML-платформой. Мониторинг.	2
16		Отладка и оптимизация ML-пайплайнов. Диагностика узких мест. Отдака сбоев. Профилирование ресурсоемких задач.	2
Всего за 5 семестр			34

3.3. Самостоятельная работа студента (СРС)

№ п/п	Номер и наименование раздела дисциплины	Содержание учебного задания	Объем, часов
1	Раздел 1. Введение в теорию больших данных.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	4
2		Подготовка к практическим занятиям	2
3	Раздел 2. Хранение и управление большими данными.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	2
4		Выполнение индивидуального практического задания	6
5		Подготовка к практическим занятиям	2
6	Раздел 3. Пакетная обработка больших данных.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	4
7		Подготовка к практическим занятиям	4
8		Выполнение индивидуального практического задания	14
9		Подготовка к контрольной работе	2
10	Раздел 4. Потокковая обработка больших данных.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	4
11		Подготовка к практическим занятиям	2
12	Раздел 5. Машинное обучение и статистический анализ больших данных.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	8
13		Подготовка к практическим занятиям	8
14		Подготовка к контрольной работе	4
15	Раздел 6. Оптимизация производительности.	Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	4
16		Подготовка к практическим занятиям	6
Всего за 5 семестр			76

4. ФОРМЫ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

СЕМЕСТР	НЕДЕЛИ СЕМЕСТРА																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
5		Вопр.Диф.Зач	Задан	ИПЗ		ДР	Задан	Контр.Р., ИПЗ		ДР	Задан	ИПЗ, Вопр.Диф.Зач			Задан	ДР	Вопр.Диф.Зач, диф. зач.

Условные обозначения:

- ДР – диагностическая работа;
- Вопр.Диф.Зач – вопросы к дифференцированному зачету;
- Задан – задание;
- ИПЗ – индивидуальное практическое задание;
- Контр.Р. – контрольная работа;
- диф. зач. – дифференцированный зачет.

Текущий контроль успеваемости студентов проводится в дискретные временные интервалы в следующих формах:

- диагностическая работа;
- вопросы к дифференцированному зачету;
- задание;
- индивидуальное практическое задание;
- контрольная работа.

Промежуточная аттестация проводится в формах:

- дифференцированный зачет.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература по дисциплине:

1. А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data. Санкт-Петербург: Лань, 2022, эл. рес.
2. Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных. М.: Юрайт, 2022, 8 экз.
3. Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных. Москва: Юрайт, 2022, эл. рес.
4. Н. А. Щипаков. . Статистические методы управления качеством. М.: Изд-во МГТУ им. Н. Э. Баумана, 2020, эл. рес.
5. С. Г. Толмачёв. . Нейросетевые методы обработки информации. СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2021, 34 экз.

5.2. Дополнительная литература по дисциплине:

не требуется.

5.3. Периодические издания:

1. Прикладная информатика.

5.4. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины, электронные библиотечные системы:

1. Федеральная служба государственной статистики - URL: <http://rosstat.gov.ru>;
2. Базы данных по курсам валют: - URL: https://cbr.ru/currency_base;
3. А.В. Макшанов. Большие данные. Big Data (2024) [Электронный ресурс] : учебник для вузов URL: <https://e.lanbook.com/book/362318>;
4. Федеральная служба государственной статистики: <http://rosstat.gov.ru>;
5. Базы данных по курсам валют: https://cbr.ru/currency_base;
6. Гидрометцентр России: <https://meteoinfo.ru>.

Современные профессиональные базы данных:

1. <https://rusneb.ru> – Национальная электронная библиотека (НЭБ);
2. <https://cyberleninka.ru/> - Научная электронная библиотека «Киберленинка»;
<http://www.rfbr.ru/rffi/ru/library> - Полнотекстовая электронная библиотека Российского фонда фундаментальных исследований.

Информационные справочные системы:

1. Техэксперт – Информационный портал технического регулирования: Нормы, правила, стандарты РФ;
2. http://library.voenmeh.ru/jirbis2/index.php?option=com_irbis&view=irbis&Itemid=457 - БД ГОСТов собственной генерации БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова;
3. <http://www.consultant.ru/>- КонсультантПлюс- информационный портал правовой информации.

5.5. Программное обеспечение:

1. Open Office.

5.6. Информационные технологии:

взаимодействие с обучающимися посредством ЭИОС Moodle БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Лекционные занятия:

специализированные требования по оборудованию отсутствуют; аудитория с посадочными местами по количеству студентов; доска.

6.2. Практические занятия:

1. Интерактивная доска;
2. Open Office.

6.3. Прочее:

1. рабочее место преподавателя, оснащенное компьютером с доступом в Интернет;
2. рабочие места студентов, оснащенные компьютерами с доступом в Интернет, предназначенные для работы в электронной образовательной среде.

Аннотация рабочей программы

Дисциплина **МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ** является дисциплиной **части, формируемой участниками образовательных отношений блока 1**, программы подготовки по направлению *09.03.04 Программная инженерия*. Дисциплина реализуется на факультете О Естественнотехнический БГТУ "ВОЕНМЕХ" им. Д.Ф. Устинова кафедрой О7 Информационные системы и программная инженерия.

Дисциплина нацелена на формирование *компетенций*:

ПК-1.4 Способен использовать технологии ИИ на различных этапах жизненного цикла программных продуктов, проектировать интеллектуальные компоненты программного обеспечения;

ПК-93 Способен генерировать новые идеи для решения задач цифровой экономики, абстрагироваться от стандартных моделей, перестраивать сложившиеся способы решения задач, выдвигать альтернативные варианты действий с целью выработки новых оптимальных алгоритмов.

Содержание дисциплины охватывает круг вопросов, связанных с использованием аналитических и графических методов для обработки больших данных; основные законы распределения случайных величин; методы сбора, обработки и анализа статистических данных в зависимости от целей исследования, техника проверки гипотез, методы корреляционного, регрессионного, кластерного и дисперсионного анализов, методы расчета показателей динамики.

Программой дисциплины предусмотрены следующие **виды контроля**:

Текущий контроль успеваемости студентов проводится в дискретные временные интервалы в следующих формах:

- диагностическая работа;
- вопросы к дифференцированному зачету;
- задание;
- индивидуальное практическое задание;
- контрольная работа.

Промежуточная аттестация проводится в формах:

- дифференцированный зачет.

Общая трудоемкость освоения дисциплины составляет 4 з.е., **144 ч**. Программой дисциплины предусмотрены лекционные занятия (**34 ч.**), практические занятия (**34 ч.**), самостоятельная работа студента (**76 ч.**).

ТЕХНОЛОГИИ И ФОРМЫ ОБУЧЕНИЯ

Рекомендации по освоению дисциплины для студента

Трудоемкость освоения дисциплины составляет 144 ч., из них 68 ч. аудиторных занятий, и 76 ч., отведенных на самостоятельную работу студента.

Рекомендации по распределению учебного времени по видам самостоятельной работы и разделам дисциплины приведены в таблице.

Контроль освоения дисциплины производится в соответствии с Положением о текущем, рубежном контроле успеваемости и промежуточной аттестации обучающихся.

Формы контроля и критерии оценивания приведены в приложении 3 к Рабочей программе.

Наименование работы	Рекомендуемая литература	Трудоемкость, час.
Раздел 1. Введение в теорию больших данных.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	С. Г. Толмачёв. . Нейросетевые методы обработки информации: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2021 (1)	4
Подготовка к практическим занятиям	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (1,2)	2
Итого по разделу 1		6
Раздел 2. Хранение и и управление большими данными.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (2,3)	2
Выполнение индивидуального практического задания		6
Подготовка к практическим занятиям		2
Итого по разделу 2		10
Раздел 3. Пакетная обработка больших данных.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных: М.: Юрайт, 2022 (3,4)	4
Подготовка к практическим занятиям	Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных: Москва: Юрайт, 2022 (3,4)	4
Выполнение индивидуального практического задания	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (3,4)	14
Подготовка к контрольной работе		2
Итого по разделу 3		24
Раздел 4. Поточковая обработка больших данных.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (6)	4
Подготовка к практическим занятиям		2
Итого по разделу 4		6
Раздел 5. Машинное обучение и статистический анализ больших данных.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкарь. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (5)	8
Подготовка к практическим занятиям	Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных: М.: Юрайт, 2022 (5,6)	8
Подготовка к контрольной работе	Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. . Статистические методы контроля качества и обработка экспериментальных данных: Москва: Юрайт, 2022 (5,6)	4

	Н. А. Щипаков. . Статистические методы управления качеством: М.: Изд-во МГТУ им. Н. Э. Баумана, 2020 (1-3)	
Итого по разделу 5		20
Раздел 6. Оптимизация производительности.		
Изучение предусмотренных программой дидактических единиц по рекомендуемой литературе	А. В. Макшанов, А. Е. Журавлёв, Л. Н. Тындыкаръ. . Большие данные. Big Data: Санкт-Петербург: Лань, 2022 (11,12, 13)	4
Подготовка к практическим занятиям		6
Итого по разделу 6		10

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств, позволяющие оценить результаты обучения по данной дисциплине, включают в себя:

- диагностическая работа
- вопросы к дифференцированному зачету;
- задание;
- индивидуальное практическое задание;
- контрольная работа;
- дифференцированный зачет.

Критерии оценивания

Диагностическая работа

Диагностическая работа проводится в форме теста в ЭИОС Moodle:

- при правильном ответе менее чем на 60% вопросов - не аттестация;
- при правильном ответе на 60% вопросов и более - аттестация.

Вопросы к дифференцированному зачету

Вопросы к дифференцированному зачету расположены в УМК дисциплины. Вопросы выдаются преподавателем заранее. При подготовке стоит пользоваться лекционным материалом, а также источниками основной и дополнительной литературы. При возникновении затруднений студент может обратиться к преподавателю в часы консультаций.

Задание

Задание представлено в срок, не представлен чужой отчет. Каждое задание разбито на 3-5 задач с последовательным увеличением нагрузки для корректного освоения требуемых компетенций. По всем заданиям необходимо успешное выполнение пунктов задания на компьютере, оформление отчета в соответствии с требованиями ГОСТ и успешная защита в установленный срок.

Количество баллов и критерии регламентируется Технологической картой дисциплины.

Индивидуальное практическое задание

Индивидуальное практическое задание выполняется на практических занятиях и в часы самостоятельной работы в соответствии с темой, определенной индивидуально для каждого обучающегося.

Практическое задание включает в себя следующие этапы:

1. Постановка цели и задач анализа больших данных.
2. Составление плана исследования в соответствии с предметной областью индивидуального задания. Формулирование гипотез для исследования.
3. Сбор данных и их группировка и систематизация.
4. Первичный анализ и описательная статистика.
5. Проверка гипотез изученными аналитическими методами; подтверждение графическими методами.
6. Оформление результатов и выводов.

Результаты выполнения этапов индивидуального практического задания выполняются средствами изученных программных пакетов и демонстрируются преподавателю на практических занятиях.

Контрольная работа

Баллы за контрольную работу проставляются согласно Технологической карте в соответствии с количеством выполненных на практическом занятии заданий средствами изученных программных пакетов.

Дифференцированный зачет

Итоговый контроль по дисциплине проходит в форме дифференцированного зачета.

Дифференцированный зачет считается сданным, если сданы все задания, в соответствии с требованиями, зафиксированными в технологической карте освоения дисциплины (не менее 51 балла).

Если обучающийся не набрал нужное количество баллов или хочет повысить оценку по дисциплине согласно технологической карте, то ему необходимо сдать Дифференцированный зачет в очном формате.

Дифференцированный зачет состоит из теоретического вопроса и практической задачи.

Критерии оценивания на Дифференцированном зачете .

Оценка «отлично»

1. Предварительно (в установленные сроки) защищены все работы в соответствии с технологической картой.
2. Даны полные ответы на вопросы (точно указаны определения, формулы, студент владеет терминологией изученной дисциплины).
3. Правильно решена задача, показано умение грамотно применять полученные теоретические знания в

практических целях.

Оценка «хорошо»

1. Предварительно (в установленные сроки) защищены работы все работы в соответствии с технологической картой.
2. Данные ответы на вопросы имеют незначительные ошибки.
3. Правильно решены задачи, но ход их решения не является оптимальным, показаны прочные практические навыки.

Оценка «удовлетворительно»

1. Работы в соответствии с технологической картой защищались с нарушением сроков сдачи.
2. Данные ответы на вопросы имеют незначительные ошибки (обучающийся в полной мере не владеет терминологией изученной дисциплины).
3. В решении задачи допущены ошибки, которые не приводят к большим отклонениям от правильного ответа, показаны не достаточно прочные практические навыки.

Оценка «неудовлетворительно»

1. Предварительно не защищены все работы в соответствии с технологической картой.
2. Ответы на вопросы имеют значительные ошибки (неточно указана формула, обучающийся не владеет терминологией изученной дисциплины).
3. Задача решена неверно, допущены грубые ошибки.

Паспорт фонда оценочных средств

КУРС	СЕМЕСТР	Наименование разделов и дидактических единиц	ВСЕГО	Аудиторные занятия в контактной форме			Самостоятельная работа студентов	Формируемая компетенция, %		НАИМЕНОВАНИЕ ОЦЕНОЧНОГО СРЕДСТВА
				ВСЕГО	Лекции	Практические занятия		ПК-1.4	ПК-93	
3	5	Раздел 1. Введение в теорию больших данных.	12	6	4	2	6	10	20	Вопросы к дифференцированному зачету
3	5	Раздел 2. Хранение и и управление большими данными.	16	6	4	2	10	20	20	Индивидуальное практическое задание, Задание
3	5	Раздел 3. Пакетная обработка больших данных.	42	18	8	10	24	20	10	Индивидуальное практическое задание, Контрольная работа, Задание
3	5	Раздел 4. Потокковая обработка больших данных.	18	12	6	6	6	20	20	Вопросы к дифференцированному зачету, Задание
3	5	Раздел 5. Машинное обучение и статистический анализ больших данных.	38	18	8	10	20	20	10	Вопросы к дифференцированному зачету, Контрольная работа, Задание
3	5	Раздел 6. Оптимизация производительности.	18	8	4	4	10	10	20	Вопросы к дифференцированному зачету, Задание
Всего за 5 семестр			144	68	34	34	76	100	100	
Всего по дисциплине			144	68	34	34	76	100	100	

Оценочные материалы по дисциплине МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

ПК-1.4 - Способен использовать технологии ИИ на различных этапах жизненного цикла программных продуктов, проектировать интеллектуальные компоненты программного обеспечения

№ 1 Прочитайте текст и запишите развернутый обоснованный ответ

Рассчитайте моду, медиану, выборочное среднее и среднеквадратическое отклонение для выборки:

3; 5; 7; 2; 5; 5

Значение среднеквадратического отклонения округлите до 1 десятичного знака.

№ 2 Прочитайте текст и запишите развернутый обоснованный ответ

Среднее по выборке измерений признака составило 23,5.

Среднеквадратическое отклонение равно 0,15.

Необходимо проверить один сомнительный результат измерения, равный 24,4.

Выберите критерий проверки, не требующий использование вспомогательных таблиц, рассчитайте необходимое значение критерия и сделайте вывод.

№ 3 Прочитайте текст и установите соответствие

Выберите метод корреляционного анализа в зависимости от типа признака-фактора и признака-результата

- | | |
|---|------------------------------------|
| 1. Оба признака количественные, непрерывные, распределены по нормальному закону | А. Коэффициент корреляции Спирмена |
| 2. Признак-фактор - качественный, может принимать 3 варианта значений. | Б. Коэффициент контингенции |
| Признак-результат непрерывный, распределен по нормальному закону | |
| 3. Оба признака альтернативные (могут принимать только 2 значения) | В. Коэффициент корреляции Пирсона |
| 4. Оба признака количественные, дискретные | Г. Коэффициент Паше |
| | Д. Коэффициент Фишера |

№ 4 Прочитайте текст и установите соответствие

Распределите методы статистического анализа в зависимости от поставленной задачи:

- | | |
|--------------------------|--|
| 1. Корреляционный анализ | А. Распределение данных на группы |
| 2. Регрессионный анализ | Б. Оценка силы связи между признаками |
| 3. Дисперсионный анализ | В. Сравнение выборок по одному или нескольким признакам |
| 4. Кластерный анализ | Г. Прогнозирование значения признака-результата по значению признака-фактора |
| | Д. Уменьшение количества признаков за счет их объединения |

№ 5 Прочитайте текст и установите последовательность

Упорядочите этапы обработки больших данных:

1. А. Анализ данных
2. Б. Постановка цели
3. В. Преобразование в единый формат
4. Г. Сбор данных
5. Д. Визуализация результатов
6. Е. Очистка данных
7. Ж. Анализ результатов

№ 6 Прочитайте текст и установите последовательность

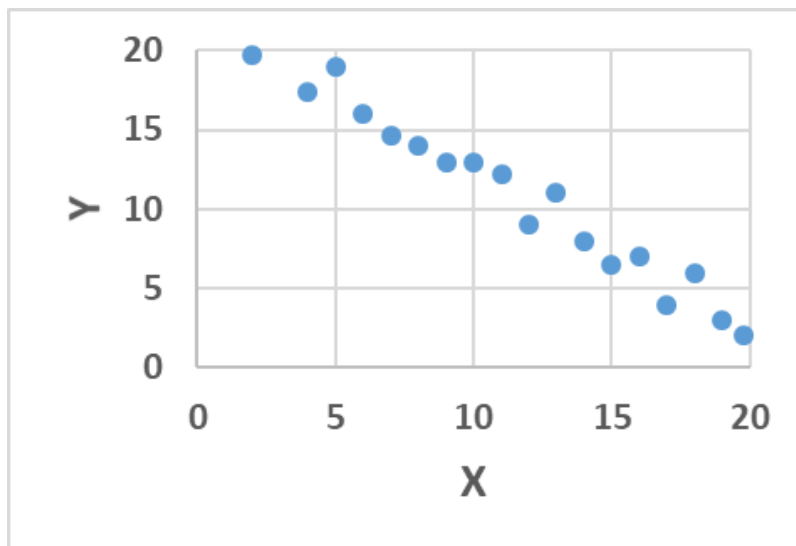
Установите последовательность расчета доверительного интервала:

1. А. Расчет стандартного отклонения по выборке
2. Б. Расчет среднего по выборке
3. В. Расчет нижней и верхней границ доверительного интервала
4. Г. Определить значение критерия Стьюдента с учетом желаемой точности оценки и размера выборки
5. Д. Расчет стандартной ошибки среднего

№ 7 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа
Разница между показателями генеральной совокупности и соответствующими показателями выборки называется:

1. Ошибкой 1-го рода
2. Ошибкой 2-го рода
3. Ошибкой репрезентативности
4. Ошибкой матожидания

№ 8 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов
Отметьте характеристики корреляционной зависимости между признаками Y и X, представленной на диаграмме рассеяния:



1. Прямая
2. Обратная
3. Сильная
4. Слабая

№ 9 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа
Определить какой показатель представленного ряда больше:

7, 3, 4, 11, 5, 5, 5, 6, 6, 5.

1. Мода
2. Медиана

3. Среднее арифметическое

4. Все эти значения равны

№ 10 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа
В результате корреляционного анализа был рассчитан коэффициент корреляции Пирсона:

$r = -0,44$.

Оцените силу связи между исследуемыми признаками по шкале Чеддока.:

1. Сильная прямая зависимость
2. Сильная обратная зависимость
3. Слабая прямая зависимость
4. Слабая обратная зависимость

№ 11 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов
Отметьте показатели вариации:

1. Стандартное отклонение
2. Размах
3. Медиана
4. Дисперсия

№ 12 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов
Отметьте типы контрольных карт Шухарта для количественных признаков:

1. R-карта
2. p-карта
3. X-карта
4. np-карта

ПК-93 - Способен генерировать новые идеи для решения задач цифровой экономики, абстрагироваться от стандартных моделей, перестраивать сложившиеся способы решения задач, выдвигать альтернативные варианты действий с целью выработки новых оптимальных алгоритмов

№ 1 Прочитайте текст и запишите развернутый обоснованный ответ
Опишите результат выполнения SQL-запроса к таблице tweets, в которой хранятся темы, упоминаемые в соцсетях.

```
SELECT themes,
COUNT(*) AS frequency
FROM tweets
WHERE date > '2025-06-01'
GROUP BY word
ORDER BY frequency DESC
LIMIT 10
```

№ 2 Прочитайте текст и запишите развернутый обоснованный ответ
В эксперименте измерения проводились 5-ью сериями. Внутригрупповая дисперсия составила 0,5, а межгрупповая - 0,01. Определите расчетное значение критерия Фишера.

№ 3 Прочитайте текст и установите соответствие
Сопоставьте термин с примером реализации.

- | | |
|-----------------------------------|--|
| 1. Горизонтальное масштабирование | А. Семантика доставки, где каждая запись обрабатывается только один раз |
| 2. CAP-теория | Б. Добавление серверов для обработки 100+ТБ данных |
| 3. РТК Хранилище | В. Российский аналог HDFS с репликацией на 3 узла |
| 4. Exactly-once | Г. Установление защиты на сервера БД |
| | Д. Гарантирует только 2 из 3-х свойств (согласованность, доступность, устойчивость к фрагментации) |

№ 4 Прочитайте текст и установите соответствие

Поставьте в соответствие проблемам представленные решения:

- | | |
|------------------------------------|--|
| 1. Утечка памяти в Pandan | А. Уменьшение Max_bin в LightGBM |
| 2. Зависание GPU при обучении | Б. Подбор гиперпараметров+конструирование признаков |
| 3. Потеря данных в Spark Streaming | В. Обработка данных чанками по 10 000 строк |
| 4. Низкое качество ML-модели | Г. Включение Write-Ahead Log |
| | Д. Конструирование признаков без подбора гиперпараметров |

№ 5 Прочитайте текст и установите последовательность

Упорядочите этапы обработки данных о ВРП регионов с использованием российского ПО.

- | | |
|---|--|
| 1 | А. Загрузка очищенных данных в аналитическую витрину ClickHouse |
| 2 | Б. Очистка данных от аномалий в Mags (фильтрация отрицательных значений) |
| 3 | В. Оркестрация пайплайна через РТК Интеграцию (ежедневный запуск в 6:00) |
| 4 | Г. Извлечение сырых CSV-файлов с FTP Росстата |
| 5 | Д. Преобразование формата дат и нормализация названий регионов |

№ 6 Прочитайте текст и установите последовательность

Восстановите порядок операций для системы мониторинга температуры с гарантией *exactly-once*.

- | | |
|---|--|
| 1 | А. Агрегация показаний по 10-секундным окнам в Spark Streaming |
| 2 | Б. Визуализация динамики в Apache Superset через подключение к ClickHouse |
| 3 | В. Запись агрегированных данных в ClickHouse |
| 4 | Г. Генерация JSON-сообщений датчиками в топик Kafka «sensors_temp» |
| 5 | Д. Настройка контрольных точек (checkpoints) в HDFS для восстановления состояния |

№ 7 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа

Выберите биологическое направление ИИ, которое используется в стиральных машинах:

1. Нейронные сети
2. Генетические алгоритмы
3. Алгоритм нечеткой логики
4. Имуннокомпьютинг
5. Эволюционное моделирование

№ 8 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов

Выберите ключевые принципы в работе с большими данными:

1. Горизонтальная адаптивность
2. Вертикальная адаптивность
3. Концентрация данных
4. Децентрализация данных
5. Стабильность в работе при отказах

№ 9 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов

Выберите критерии информации, которые позволяют оценить, соответствуют ли данные понятию Big Data или нет:

1. Volume (объем)
2. Velocity (скорость)
3. Validity (валидность)
4. Virtuality (виртуальность)
5. Variety (разнообразие)

№ 10 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов
Отметьте основные задачи машинного обучения:

1. Регрессия
2. Классификация
3. Факторизация
4. Кластеризация
5. Поиск аномалий

№ 11 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа
При обучении модели LightGBM на Платформе O7 перегревается GPU, время выполнения превышает 4 часа.

Выберите наиболее эффективное решение для оптимизации?

1. Увеличение размера микропакетов (batch size) в 2 раза.
2. Уменьшение параметра `max_bin` для снижения нагрузки.
3. Добавление 10 новых слоёв в нейронную сеть.
4. Отключение контрольных точек (checkpoints) в РТК Интеграции.

№ 12 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа
Нужно обрабатывать 10 млрд строк данных Росстата с агрегациями (медианы, квантили) в реальном времени. Выберите оптимальное для этой задачи российское ПО из реестра:

1. Тарпан Хранилище (документная NoSQL)
2. ClickHouse (колоночная СУБД)
3. Postgres Pro (реляционная СУБД)
4. РТК Хранилище (распределённая ФС)