МИНОБРНАУКИ РОССИИ

ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Рабочая программа дисциплины (модуля)

ВВЕДЕНИЕ В АНАЛИЗ БОЛЬШИХ ДАННЫХ

Специальность и специализация
10.05.03 Информационная безопасность автоматизированных систем. Безопасность открытых информационных систем

Год набора на ОПОП
2022

Форма обучения
очная

Владивосток 2026

Рабочая программа дисциплины (модуля) «Введение в анализ больших данных» составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 10.05.03 Информационная безопасность автоматизированных систем (утв. приказом Минобрнауки России от 26.11.2020г. №1457) и Порядком организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры (утв. приказом Минобрнауки России от 06.04.2021 г. N245).

Составитель(и):

Ермолицкая М.З.

Утверждена на заседании научно-образовательный центр "искусственный интеллект" от 27.05.2026 , протокол № 5

СОГЛАСОВАНО:

Заведующий кафедрой (разработчика)

Кригер А.Б.

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат	1582918206
Номер транзакции	0000000000F9A70F
Владелец	Кригер А.Б.

1 Цель, планируемые результаты обучения по дисциплине (модулю)

Целью освоения дисциплин ы «Введение в анализ больших данных» является теоретическая и практическая подготовка студентов к работе с большими данными. Знания, полученные в результате освоения дисциплины, помогут при сборе и анализе огромных объемов структурированной или неструктурированной информации, при разработке моделей данных и получении новых знаний. Все это необходимо выпускнику для решения различных задач практической и научно-исследовательской деятельности.

Задачи освоения дисциплины:

· приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;

· применение статистических и математических методов для анализа больших объемов информации;

· приобретение практических навыков работы с программой RStudio.

Планируемыми результатами обучения по дисциплине (модулю), являются знания, умения, навыки. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы, представлен в таблице 1.

Таблица 1 – Компетенции, формируемые в результате изучения дисциплины (модуля)

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции	Результаты обучения по дисциплине		
			Код результата	Формулировка результата	
10.05.03 «Информационная безопасность автоматизированных систем» (ИБ)	ОПК-1 : Способен оценивать роль информации, информационных технологий и информационной безопасности в современном обществе, их значение для обеспечения объективных потребностей личности, общества и государства	ОПК-1.1к : Понимает принципы теории информационной безопасности и проблемы государственной и региональной информационной безопасности	РД1	Знание	основных методов обработки и анализа больших данных
			РД2	Умение	проводить сравнительный анализ и выбор статистических методов для анализа конкретных данных
			РД3	Навык	применения статистических методов для обработки и анализа больших объемов информации с использованием программы RStudio

В процессе освоения дисциплины решаются задачи воспитания гармонично развитой, патриотичной и социально ответственной личности на основе традиционных российских духовно-нравственных и культурно-исторических ценностей, представленные в таблице 1.2.

Таблица 1.2 – Целевые ориентиры воспитания

Воспитательные задачи	**Формирование ценностей**	**Целевые ориентиры**
Формирование научного мировоззрения и культуры мышления		

Формирование осознания ценности научного мировоззрения и критического мышления	Гуманизм	Системное мышление
Формирование коммуникативных навыков и культуры общения		
Формирование навыков публичного выступления и презентации своих идей	Взаимопомощь и взаимоуважение	Умение работать в команде и взаимопомощь

2 Место дисциплины (модуля) в структуре ОПОП

Дисциплина относится к обязательной части учебного плана Блоку 1 ("Дисциплины. модули")

3. Объем дисциплины (модуля)

Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу с обучающимися (по видам учебных занятий) и на самостоятельную работу, приведен в таблице 2.

Таблица 2 – Общая трудоемкость дисциплины

Название ОПОП ВО	Форма обучения	Часть УП	Семестр (ОФО) или курс (ЗФО, ОЗФО)	Трудоемкость	Объем контактной работы (час)						СРС	Форма аттестации
				(З.Е.)	Всего	Аудиторная			Внеаудиторная			
						лек.	прак.	лаб.	ПА	КСР		
10.05.03 Информационная безопасность автоматизированных систем	ОФО	С4.Ф	11	3	37	12	24	0	1	0	71	Э

4 Структура и содержание дисциплины (модуля)

4.1 Структура дисциплины (модуля) для ОФО

Тематический план, отражающий содержание дисциплины (перечень разделов и тем), структурированное по видам учебных занятий с указанием их объемов в соответствии с учебным планом, приведен в таблице 3.1

Таблица 3.1 – Разделы дисциплины (модуля), виды учебной деятельности и формы текущего контроля для ОФО

№	Название темы	Код результата обучения	Кол-во часов, отведенное на				Форма текущего контроля
			Лек	Практ	Лаб	СРС	
1	Введение в анализ больших данных. Обзор источников информации.	РД1	2	0	0	10	собеседование
2	Технологии хранения и обработки больших данных.	РД1	2	0	0	6	собеседование
3	Современные программные средства анализа больших объемов информации.	РД1	2	0	0	6	собеседование

4	Методы обработки и анализа больших данных.	РД1	10	0	0	10	собеседование
5	Сбор и хранение больших данных.	РД2	0	6	0	10	отчет по выполненным практическим работам
6	Методы обработки, анализа и визуализации больших данных в программе RStudio.	РД2, РД3	0	26	0	17	отчет по выполненным практическим работам
Итого по таблице			**16**	**32**	**0**	**59**	

4.2 Содержание разделов и тем дисциплины (модуля) для ОФО

Тема 1 Введение в анализ больших данных. Обзор источников информации.

Содержание темы: Основные определения, термины, задачи анализа больших данных. Вопросы безопасности. Понятие Data Mining. Когнитивный анализ данных. Обзор источников информации для Big Data (открытые источники информации: статистические сборники, опубликованные отчеты и результаты исследований; доступ к закрытой информации). Методики сбора данных.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: лекции, на которых дается основной систематизированный материал по темам.

Виды самостоятельной подготовки студентов по теме: чтение предлагаемой литературы, подготовка к собеседованию, итоговому тесту.

Тема 2 Технологии хранения и обработки больших данных.

Содержание темы: Обзор технологий хранения больших данных. Базы данных. Системы управления базами данных. Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: лекции, на которых дается основной систематизированный материал по темам.

Виды самостоятельной подготовки студентов по теме: чтение предлагаемой литературы, подготовка к собеседованию, итоговому тесту.

Тема 3 Современные программные средства анализа больших объемов информации.

Содержание темы: Обзор современных популярных программных средств анализа данных. Платные, бесплатные программные средства: Statistica, SPSS, Excel, R-Studio и другие; их преимущества и недостатки.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: лекции, на которых дается основной систематизированный материал по темам.

Виды самостоятельной подготовки студентов по теме: чтение предлагаемой литературы, подготовка к собеседованию, итоговому тесту.

Тема 4 Методы обработки и анализа больших данных.

Содержание темы: Основные понятия математической статистики. Методы анализа данных: дескриптивная статистика, критерии для проверки на нормальность распределения; параметрические, непараметрические, номинальные методы (критерии для определения значимости различий в выборках, определение зависимости между переменными, построение регрессионных моделей, дисперсионный, кластерный, дискриминантный, факторный анализы).

Формы и методы проведения занятий по теме, применяемые образовательные технологии: лекции, на которых дается основной систематизированный материал по темам.

Виды самостоятельной подготовки студентов по теме: чтение предлагаемой литературы, подготовка к собеседованию, итоговому тесту.

Тема 5 Сбор и хранение больших данных.

Содержание темы: Поиск источников информации в сети Интернет: открытые и закрытые источники данных. Портал открытых данных РФ. Сохранение данных в программе MS Excel. Преобразование и первичная обработка данных.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: практические занятия проводятся в компьютерном классе с использованием программ RStudio (преподаватель излагает тему, приводит примеры и дает задание для самостоятельного выполнения, при необходимости консультирует студентов).

Виды самостоятельной подготовки студентов по теме: подготовка отчета по практическим работам, подготовка к итоговому тесту.

Тема 6 Методы обработки, анализа и визуализации больших данных в программе RStudio.

Содержание темы: Представление исходных данных в программе RStudio (векторы, массивы, матрицы, списки, таблицы). Статистическая обработка данных в программах MS Excel и RStudio: подсчет описательных статистик, графическое представление данных. Группировка данных, обнаружение значимых зависимостей и тенденций в результате анализа имеющейся информации, выявления отношений между данными различного типа. Применение различных методов выделения, извлечения и группировки данных, которые позволяют выявить систематизированные структуры данных и вывести из них правила для принятия решений и прогнозирования их последствий (регрессионный, дисперсионный, кластерный, дискриминантный, факторный анализы). Возможности графического представления информации в программе RStudio: графические функции отображения одномерных и многомерных данных, графический вывод с использованием графических параметров.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: практические занятия проводятся в компьютерном классе с использованием программы RStudio (преподаватель излагает тему, приводит примеры и дает задание для самостоятельного выполнения, при необходимости консультирует студентов).

Виды самостоятельной подготовки студентов по теме: подготовка отчета по практическим работам, подготовка к итоговому тесту.

5 Методические указания для обучающихся по изучению и реализации дисциплины (модуля)

5.1 Методические рекомендации обучающимся по изучению дисциплины и по обеспечению самостоятельной работы

На лекциях дается основной систематизированный материал по темам.

Практические занятия проводятся в компьютерном классе с использованием программ MS Excel и RStudio. Преподаватель излагает тему, приводит примеры и дает задание для самостоятельного выполнения. При необходимости консультирует студентов.

Самостоятельная работа студентов подразумевает чтение предлагаемой преподавателем литературы и использование интернет-ресурсов для подготовки к занятиям, текущей и промежуточной аттестации.

Промежуточная аттестация - итоговый тест.

5.2 Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

При необходимости обучающимся из числа лиц с ограниченными возможностями здоровья и инвалидов (по заявлению обучающегося) предоставляется учебная информация в доступных формах с учетом их индивидуальных психофизических особенностей:

- для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания, консультации и др.
- для лиц с нарушениями слуха: в печатной форме; в форме электронного документа; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания, консультации и др.
- для лиц с нарушениями опорно-двигательного аппарата: в печатной форме; в форме электронного документа; индивидуальные задания, консультации и др.

6 Фонд оценочных средств для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине (модулю)

В соответствии с требованиями ФГОС ВО для аттестации обучающихся на соответствие их персональных достижений планируемым результатам обучения по дисциплине (модулю) созданы фонды оценочных средств. Типовые контрольные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 1.

7 Учебно-методическое и информационное обеспечение дисциплины (модуля)

7.1 Основная литература

1. Кремер, Н. Ш. Теория вероятностей : учебник и практикум для вузов / Н. Ш. Кремер. — 5-е изд. — Москва : Издательство Юрайт, 2025. — 259 с. — (Высшее образование). — ISBN 978-5-534-17131-0. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/561038 (дата обращения: 01.09.2025).
2. Миркин, Б. Г. Базовые методы анализа данных : учебник и практикум для вузов / Б. Г. Миркин. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2024. — 261 с. — (Высшее образование). — ISBN 978-5-534-18842-4. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/551786 (дата обращения: 12.03.2025).
3. Статистика. В 2 ч. Часть 2 : учебник и практикум для вузов / В. С. Мхитарян, Т. Н. Агапова, С. Д. Ильенкова, А. Е. Суринов ; под редакцией В. С. Мхитаряна. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2023. — 270 с. — (Высшее образование). — ISBN 978-5-534-09357-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/517262 (дата обращения: 01.03.2023).

7.2 Дополнительная литература

1. Гуриков, С. Р. Основы алгоритмизации и программирования на Python : учебное пособие / С.Р. Гуриков. — Москва : ИНФРА-М, 2025. — 343 с. — (Высшее образование). - ISBN 978-5-16-020255-6. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2166199 (дата обращения: 31.05.2026)
2. Панов, М. А. Анализ данных с использованием языка программирования Python : учебное пособие / М. А. Панов. — Екатеринбург : УрГЭУ, 2024. — 329 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/481577 (дата обращения: 25.05.2026). — Режим доступа: для авториз. пользователей.

7.3 Ресурсы информационно-телекоммуникационной сети "Интернет", включая профессиональные базы данных и информационно-справочные системы (при необходимости):

1. Образовательная платформа "ЮРАЙТ"
2. Образовательная платформа "ЮРАЙТ" - Режим доступа: https://urait.ru/
3. Электронно-библиотечная система "ZNANIUM.COM"
4. Электронно-библиотечная система "ЛАНЬ"
5. Open Academic Journals Index (OAJI). Профессиональная база данных - Режим доступа: http://oaji.net/
6. Президентская библиотека им. Б.Н.Ельцина (база данных различных профессиональных областей) - Режим доступа: https://www.prlib.ru/
7. Информационно-справочная система "Консультант Плюс" - Режим доступа: http://www.consultant.ru/

8 Материально-техническое обеспечение дисциплины (модуля) и перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

Основное оборудование:

- Компьютеры
- Монитор облачный 23" LG23CAV42K/мышь Genius Optical Wheel проводная/клавиатура Genius KB110 проводная
- Мультимедийный проектор CASIO (Япония)
- Облачный монитор LG Electronics черный +клавиатура+мышь
- Уст-во бесп.пит.SmartUPS 3000

Программное обеспечение:

- Python
- RStudio

МИНОБРНАУКИ РОССИИ

ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Фонд оценочных средств
для проведения текущего контроля
и промежуточной аттестации по дисциплине (модулю)

ВВЕДЕНИЕ В АНАЛИЗ БОЛЬШИХ ДАННЫХ

Специальность и специализация
10.05.03 Информационная безопасность автоматизированных систем. Безопасность открытых информационных систем

Год набора на ОПОП
2022

Форма обучения
очная

Владивосток 2026

1 Перечень формируемых компетенций

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции
10.05.03 «Информационная безопасность автоматизированных систем» (ИБ)	ОПК-1 : Способен оценивать роль информации, информационных технологий и информационной безопасности в современном обществе, их значение для обеспечения объективных потребностей личности, общества и государства	ОПК-1.1к : Понимает принципы теории информационной безопасности и проблемы государственной и региональной информационной безопасности

Компетенция считается сформированной на данном этапе в случае, если полученные результаты обучения по дисциплине оценены положительно (диапазон критериев оценивания результатов обучения «зачтено», «удовлетворительно», «хорошо», «отлично»). В случае отсутствия положительной оценки компетенция на данном этапе считается несформированной.

2 Показатели оценивания планируемых результатов обучения

Компетенция ОПК-1 «Способен оценивать роль информации, информационных технологий и информационной безопасности в современном обществе, их значение для обеспечения объективных потребностей личности, общества и государства»

Таблица 2.1 – Критерии оценки индикаторов достижения компетенции

Код и формулировка индикатора достижения компетенции	Результаты обучения по дисциплине			Критерии оценивания результатов обучения
	Код рез-та	Тип рез-та	Результат	
ОПК-1.1к : Понимает принципы теории информационной безопасности и проблемы государственной и региональной информационной безопасности	РД 1	Знание	основных методов обработки и анализа больших данных	знание методов обработки и анализа больших данных, используемых при решении профессиональных задач
	РД 2	Умение	проводить сравнительный анализ и выбор статистических методов для анализа конкретных данных	проведение сравнительного анализа и выбора статистических методов для анализа конкретных данных
	РД 3	Навык	применения статистических методов для обработки и анализа больших объемов информации с использованием программы RStudio	применение статистических методов для обработки и анализа больших объемов информации с использованием программы RStudio

Таблица заполняется в соответствии с разделом 1 Рабочей программы дисциплины (модуля).

3 Перечень оценочных средств

Таблица 3 – Перечень оценочных средств по дисциплине (модулю)

Контролируемые планируемые результаты обучения		Контролируемые темы дисциплины	Наименование оценочного средства и представление его в ФОС	
			Текущий контроль	Промежуточная аттестация
Очная форма обучения				
РД1	Знание : основных методов обработки и анализа больших данных	1.1. Введение в анализ больших данных. Обзор источников информации.	Тест	Тест
		1.2. Технологии хранения и обработки больших данных.	Тест	Тест
		1.3. Современные программные средства анализа больших объемов информации.	Тест	Тест
		1.4. Методы обработки и анализа больших данных.	Тест	Тест
РД2	Умение : проводить сравнительный анализ и выбор статистических методов для анализа конкретных данных	1.5. Сбор и хранение больших данных.	Собеседование	Тест
		1.6. Методы обработки, анализа и визуализации больших данных в программе RStudio.	Собеседование	Тест
РД3	Навык : применения статистических методов для обработки и анализа больших объемов информации с использованием программы RStudio	1.6. Методы обработки, анализа и визуализации больших данных в программе RStudio.	Тест	Тест

4 Описание процедуры оценивания

Качество сформированности компетенций на данном этапе оценивается по результатам текущих и промежуточных аттестаций при помощи количественной оценки, выраженной в баллах. Максимальная сумма баллов по дисциплине (модулю) равна 100 баллам.

Вид учебной деятельности	Оценочное средство			
	Собеседование	Отчет по выполненным практическим работам	Итоговый тест	Итого
Лекции	10			10
Практические занятия		60		60
Самостоятельная работа	10			10
Итоговая аттестация			20	20
Итого	20	60	20	100

Сумма баллов, набранных студентом по всем видам учебной деятельности в рамках дисциплины, переводится в оценку в соответствии с таблицей.

Сумма баллов по дисциплине	Оценка по промежуточной аттестации	Характеристика качества сформированности компетенции
от 91 до 100	«зачтено» / «отлично»	Студент демонстрирует сформированность дисциплинарных компетенций, обнаруживает всестороннее, систематическое и глубокое знание учебного материала, усвоил основную литературу и знаком с дополнительной литературой, рекоме

		ндованной программой, умеет свободно выполнять практические задания, пред усмотренные программой, свободно оперирует приобретенными знаниями, уме ниями, применяет их в ситуациях повышенной сложности.
от 76 до 90	«зачтено» / «хорошо»	Студент демонстрирует сформированность дисциплинарных компетенций: осно вные знания, умения освоены, но допускаются незначительные ошибки, неточн ости, затруднения при аналитических операциях, переносе знаний и умений на н овые, нестандартные ситуации.
от 61 до 75	«зачтено» / «удовлетворитель но»	Студент демонстрирует сформированность дисциплинарных компетенций: в хо де контрольных мероприятий допускаются значительные ошибки, проявляется о тсутствие отдельных знаний, умений, навыков по некоторым дисциплинарным к омпетенциям, студент испытывает значительные затруднения при оперировани и знаниями и умениями при их переносе на новые ситуации.
от 41 до 60	«не зачтено» / «неудовлетворите льно»	У студента не сформированы дисциплинарные компетенции, проявляется недос таточность знаний, умений, навыков.
от 0 до 40	«не зачтено» / «неудовлетворите льно»	Дисциплинарные компетенции не сформированы. Проявляется полное или прак тически полное отсутствие знаний, умений, навыков.

5 Примерные оценочные средства

5.1 Примерный перечень вопросов по темам

Контрольные вопросы для собеседования по темам

Тема 1. Введение в анализ больших данных. Обзор источников информации.

1. Дайте определение понятию «информационные ресурсы».
2. Что означает «информационный поиск»?
3. Информационно-коммуникационные технологии, что это?
4. Перечислите основные компоненты процесса поиска информации.
5. Определите понятие «информационные системы».
6. Охарактеризуйте портал открытых данных РФ.
7. Определите сущность понятия «большие данные».
8. Определите понятие Data Mining.
9. Каковы главные проблемы безопасности «больших данных»?
10. Дайте характеристику принципу безопасности «intelligence».

Тема 2. Технологии хранения и обработки больших данных.

1. Перечислите технологии хранения больших данных.
2. Характеристики системы хранения данных RCS.
3. Анализ больших данных в QlikView.
4. Характеристики системы хранения данных Полибайт.
5. Какие модели данных вы знаете?
6. Что включает первичная обработка данных?

Тема 3. Современные программные средства анализа больших объемов информации.

1. Перечислите программные средства анализа данных: платные и бесплатные.
2. Преимущества работа с данными в программе R-Studio.
3. Каковы возможности представления данных в программе R-Studio?

Тема 4. Статистические методы анализа данных.

1. Опишите свойства нормального распределения.

2. Определите различия между параметрическими, непараметрическими и номинальными методами.
3. Критерии для определения различий в выборках.
4. Опишите основную идею корреляционного анализа.
5. Что показывает коэффициент корреляции Пирсона?
6. Коэффициенты связи между переменными, не подчиняющимися нормальному закону распределения.
7. Для чего применяют регрессионный анализ?
8. Типы регрессионных моделей.
9. Как проверить адекватность построенной регрессионной модели?
10. Основная идея дисперсионного анализа.
11. Сущность кластерного анализа.
12. Для чего используют дискриминантный анализ?
13. Цели применения факторного анализа.

Краткие методические указания

Собеседование проводится после изучения соответствующей темы. Преподаватель в устной форме задает вопросы студентам на лекционных занятиях.

Шкала оценки

№	Баллы	Описание
5	16–20	Процент правильных и обоснованных ответов от 95% до 100%
4	11–15	Процент правильных и обоснованных ответов от 80 до 94%
3	6–10	Процент правильных ответов с помощью наводящих вопросов от 65 до 79%
2	0–5	Процент правильных ответов от 45 до 64%

5.2 Примеры заданий для выполнения практических работ

Примерный перечень тематики практических заданий, по результатам которых проводится тестирование умений и навыков:

Тема 1. Сбор данных из различных источников в сети Интернет. Портал открытых данных РФ. Хранение данных в программе MS Excel.

Тема 2. Первичный анализ данных и визуализация данных в программе Excel.

Тема 3. Знакомство с программой RStudio. Синтаксис. Представление исходных данных в программе RStudio (векторы, массивы, матрицы, списки, таблицы).

Тема 4. Выборка и преобразование исходных данных в программе RStudio. Удаление пропущенных значений.

Тема 5. Статистическая обработка данных в программе RStudio: подсчет описательных статистик.

Тема 6. Возможности графического представления информации в программе RStudio: графические функции отображения одномерных и многомерных данных, графический вывод с использованием графических параметров.

Тема 7. Законы распределения вероятностей, реализованные в R. Проверка данных на нормальность распределения: критерии Шапиро-Уилка, Колмогорова-Смирнова и др. Уровень статистической достоверности.

Тема 8. Сравнение выборок. Критерий Стьюдента, Критерий согласия хи-квадрат Пирсона, Критерий Колмогорова-Смирнова.

Тема 9. Непараметрические методы сравнения для зависимых и независимых выборок: критерий Уилкоксона, критерий Краскела-Уоллиса.

Тема 10. Корреляционный анализ. Расчет коэффициентов корреляции Пирсона, Спирмена, Кендалла

Тема 11. Регрессионный анализ (линейная зависимость). Построение линейной модели. Проверка адекватности построенной модели.

Тема 12. Регрессионный анализ (нелинейная зависимость). Определение вида зависимости. Построение модели. Проверка адекватности построенной модели.

Тема 13. Однофакторный дисперсионный анализ.

Тема 14. Многофакторный дисперсионный анализ.

Тема 15. Факторный анализ.

Тема 16. Кластерный анализ.

Тема 17. Возможности графического представления информации в программе RStudio: графические функции отображения одномерных и многомерных данных, графический вывод с использованием графических параметров.

Краткие методические указания

На выполнение одной практической работы отводится не более двух академических часов. Темы логически объединяются в Модули (3-4 темы на один модуль).

После выполнения каждого учебного модуля студент выполняет проверочный тест. В тест включены как концептуальные вопросы, так и вопросы, связанные с созданием скрипта, с оценкой вычисляемых показателей.

Шкала оценки

№	Баллы	Описание
5	47–60	Студент демонстрирует умения на итоговом уровне: умеет свободно выполнять практические задания, предусмотренные программой, свободно оперирует приобретенными знаниями и умениями, применяет их в ситуациях повышенной сложности.
4	32–46	Студент демонстрирует умения на среднем уровне: освоил основные умения, но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях, переносе умений на новые, нестандартные ситуации.
3	26–31	Студент демонстрирует умения и навыки на базовом уровне: в ходе контрольных мероприятий допускаются значительные ошибки, проявляется отсутствие отдельных умений, навыков по дисциплинарным компетенциям, испытываются значительные затруднения при оперировании умениями и при их переносе на новые ситуации.
2	0–25	Студент демонстрирует умения и навыки на уровне ниже базового: проявляется недостаточность умений и навыков.

5.3 Итоговый тест

Данные - это

a) факты, характеризующие объекты, процессы, явления предметной области

b) данные, рассматриваемые в каком-либо контексте, из которого пользователь может составить собственное мнение

c) закономерности проблемной области, полученные в результате практической деятельности и профессионального опыта, позволяющие специалистам ставить и решать задачи в этой области

d) сведения, передаваемые людьми устным, письменным или другим способом

Каким признаком не обладают большие данные?

a) многообразие данных;

b) достоверность данных;

c) скорость накопления данных;

d) типизацией исходных данных.

Слабо структурированные данные могут быть записаны в форме

a) таблиц;

b) отдельных векторов (строк);

c) записей произвольной последовательности;

d) таблиц и отношений.

Неструктурированные данные могут быть записаны в форме

a) таблиц;

b) записей произвольной последовательности;

c) таблиц и отношений;

d) записей произвольной длины.

Аналитик- это

a) специалист, занимающийся анализом в различных сферах деятельности и разработкой моделей ля проведения анализа;

b) специалист в выбранной предметной области;

c) сотрудник выполняющих узкий круг задач в выбранной области;

d) сотрудник, который имеет опыт в программировании.

Эксперт - это

a) специалист, занимающийся анализом в различных областях деятельности и разработкой моделей ля проведения анализа;

b) специалист в выбранной предметной области;

c) сотрудник выполняющих узкий круг задач в выбранной области;

d) сотрудник, который имеет опыт в программировании.

Классификация -

a) некоторый набор операций над базой данных, который рассматривается, как единственное завершено, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;

b) разновидность систем хранения, ориентирована на поддержку процесса анализа данных и их целостность;

c) высокоуровневые средства отражения информационной модели и описания структуры данных;

d) это установление зависимости дискретной выходной переменной от входных переменных.

Обучающая выборка -

a) эта группировка объектов (наблюдений) на основе данных, описывающих свойства объектов;

b) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданные входы, и соответствующий правильный выходной результат;

c) выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Ошибка обучения -

a) это ошибка, допущенная моделью на учебном множестве;

b) это ошибка, полученная на тестовых примерах;

c) имена, типы и назначения полей исходной выборки данных;

d) набор данных, каждая запись которого представляет собой учебный пример, содержащего заданные входы, и соответствующий правильный выходной результат.

Данные, рассматриваемые в каком-либо контексте, из которого пользователь может составить собственное мнение - это

a) данные;

b) знания;

c) информация.

Множество примеров, используемое для проверки работы сконструированной модели, называется

a) тестовым множеством;

b) множеством входных переменных;

c) обучающим множеством;

d) множеством выходных переменных.

Модель (алгоритм) называют обучаемым если

a) модель осуществляет интерактивное взаимодействие с экспертом;

b) модель (алгоритм) самостоятельно обнаруживает в данных присутствующие в них закономерности;

c) модель (алгоритм) самостоятельно использует известные закономерности.

Таблицы в базах данных предназначены для

a) хранения данных базы;

b) отбора и обработки данных базы;

c) автоматического выполнения группы команд;

d) визуализации данных.

Data Mining включает в себя

a) один базовый метод обнаружения знаний;

b) только статистические методы обработки данных для извлечения знаний;

c) большое число различных методов извлечения знаний;

d) большое число статистических методов извлечения знаний.

Целью построения модели регрессии можно назвать

a) прогнозирование числовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных;

b) объединение объектов или наблюдений, на основе близости значений их атрибутов (признаков);

c) исследование взаимной связи между объектами и/или событиями;

d) разбиения множества объектов или наблюдений на априорно заданные группы.

Целью задачи классификации можно назвать

a) числовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных;

b) объединение объектов или наблюдений, на основе близости значений их атрибутов (признаков);

c) исследование взаимной связи между объектами и/или событиями;

d) разбиения множества объектов или наблюдений на априорно заданные группы.

Целью задачи кластеризации можно назвать

a) числовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных;

b) объединение объектов или наблюдений, на основе близости значений их атрибутов (признаков);

c) исследование взаимной связи между объектами и/или событиями;

d) разбиения множества объектов или наблюдений на априорно заданные группы.

Целью формирования ассоциативных правил можно назвать

a) числовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных;

b) объединение объектов или наблюдений, на основе близости значений их атрибутов (признаков);

c) установление взаимной связи между объектами и/или событиями;

d) разбиения множества объектов или наблюдений на априорно заданные группы.

К классу прогнозирующих задач Data Mining относится

a) кластеризация;

b) поиск ассоциативных правил;

c) регрессия;

d) Классификация.

Два основных типа данных в статистике

a) качественные и количественные;

b) количественные и символьные;

c) текстовые и числовые;

d) векторы и массивы.

Краткие методические указания

Итоговый тест проводится в электронной форме во время последнего в учебном периоде практического занятия. Тест состоит из 20 тестовых заданий. На выполнение теста отводится 20 минут. Во время проведения теста использование литературы и других информационных ресурсов допускается только по предварительному согласованию с преподавателем.

Шкала оценки

№	Баллы	Описание
5	16–20	Процент правильных ответов от 91% до 100%
4	11–15	Процент правильных ответов от 80 до 90%
3	6–10	Процент правильных ответов от 65 до 79%
2	0–5	Процент правильных ответов от 40 до 64%

ФОС и ключи для ФОС
по дисциплине «Введение в анализ больших данных»

5.1 Вопросы и ответы

Тема 1. Введение в анализ больших данных. Обзор источников информации.

1. Информационные ресурсы — это организованные совокупности данных, документов, знаний и других информационных объектов, которые используются для удовлетворения информационных потребностей пользователей.

2. Информационный поиск — процесс нахождения информации в различных источниках, соответствующих запросу пользователя.

3. Информационно-коммуникационные технологии (ИКТ) — это технологии, используемые для обработки, хранения, передачи и распространения информации, включая компьютеры, сети, программное обеспечение и телекоммуникации.

4. Основные компоненты процесса поиска информации:
- Формулировка запроса.
- Выбор источников информации.
- Поиск и извлечение данных.
- Оценка релевантности результатов.
- Представление информации пользователю.

5. Информационные системы — это системы, предназначенные для сбора, хранения, обработки, поиска и распространения информации.

6. Портал открытых данных РФ — это платформа, предоставляющая доступ к открытым данным государственных органов России для свободного использования.

7. Большие данные (Big Data) — это огромные объемы структурированных и неструктурированных данных, которые сложно обрабатывать традиционными методами.

8. Data Mining — процесс анализа больших объемов данных для выявления скрытых закономерностей, тенденций и взаимосвязей.

9. Главные проблемы безопасности больших данных:
- Утечки данных.
- Несанкционированный доступ.
- Сложность обеспечения конфиденциальности.
- Соответствие законодательству (например, GDPR).

10. Принцип безопасности "intelligence" — это использование аналитических методов для прогнозирования и предотвращения угроз безопасности данных.

Тема 2. Технологии хранения и обработки больших данных.

1. Технологии хранения больших данных:
   - Hadoop (HDFS).
   - NoSQL (MongoDB, Cassandra).
   - SQL-базы данных (PostgreSQL, MySQL).
   - Облачные хранилища (Amazon S3, Google Cloud Storage).

2. Характеристики системы хранения данных RCS:
   - Высокая масштабируемость.
   - Поддержка распределенного хранения.
   - Отказоустойчивость.

3. Анализ больших данных в QlikView:
   - Визуализация данных.
   - Интерактивные дашборды.
   - Поддержка интеграции с различными источниками данных.

4. Характеристики системы хранения данных Полибайт:
   - Поддержка больших объемов данных.
   - Высокая скорость обработки.
   - Гибкость в настройке.

5. Модели данных:
   - Реляционная.
   - Иерархическая.
   - Сетевая.
   - Объектно-ориентированная.
   - Документоориентированная (NoSQL).

6. Первичная обработка данных включает:
   - Очистку данных.
   - Нормализацию.
   - Удаление дубликатов.
   - Преобразование форматов.

Тема 3. Современные программные средства анализа больших объемов информации.

1. Программные средства анализа данных:
   - Платные: Tableau, SAS, IBM SPSS, MATLAB.
   - Бесплатные: R, Python (Pandas, NumPy), KNIME, Weka.

2. Преимущества работы с данными в R-Studio:
   - Широкий набор статистических методов.
   - Гибкость в визуализации данных.

- Большое сообщество пользователей.

3. Возможности представления данных в R-Studio:
- Графики (ggplot2).
- Интерактивные дашборды (Shiny).
- Таблицы и отчеты.

Тема 4. Статистические методы анализа данных.

1. Свойства нормального распределения:
- Симметричность.
- Унимодальность (один пик).
- Среднее, мода и медиана совпадают.

2. Различия между методами:
- Параметрические: предполагают нормальное распределение.
- Непараметрические: не требуют предположений о распределении.
- Номинальные: работают с категориальными данными.

3. Критерии для определения различий в выборках:
- t-критерий (параметрический).
- U-критерий Манна-Уитни (непараметрический).

4. Корреляционный анализ — изучение взаимосвязи между переменными.

5. Коэффициент корреляции Пирсона показывает силу и направление линейной связи.

6. Коэффициенты связи для ненормальных распределений:
- Коэффициент Спирмена.
- Коэффициент Кендалла.

7. Регрессионный анализ применяется для прогнозирования и моделирования зависимостей.

8. Типы регрессионных моделей:
- Линейная.
- Логистическая.
- Полиномиальная.

9. Проверка адекватности регрессионной модели:
- Анализ остатков.
- Проверка на мультиколлинеарность.
- Использование критериев (R^2, F-критерий).

10. Дисперсионный анализ (ANOVA) — сравнение средних значений в группах.

11. Кластерный анализ — группировка объектов по схожести.

12. Дискриминантный анализ — классификация объектов в группы.

13. Факторный анализ — сокращение числа переменных и выявление скрытых факторов.

Дискриминантный анализ (детализация)

1. Определение
Дискриминантный анализ (ДА) — это статистический метод, предназначенный для классификации объектов в заранее определённые группы на основе набора предикторных переменных (признаков). Он также помогает понять, какие переменные наиболее значимы для разделения групп.

2. Основные цели
- Классификация: Отнесение новых наблюдений к одной из существующих групп (например, определение, является ли клиент надёжным заёмщиком).
- Анализ различий: Выявление признаков, которые наиболее сильно различают группы (например, какие факторы влияют на выбор марки автомобиля).
- Сокращение размерности: Проецирование данных в пространство меньшей размерности (аналогично методу главных компонент, но с учётом информации о группах).

3. Типы дискриминантного анализа
- Линейный дискриминантный анализ (LDA): Используется, когда ковариационные матрицы групп одинаковы.
Пример: Разделение пациентов на группы "здоровые" и "больные" на основе анализов крови.
- Квадратичный дискриминантный анализ (QDA): Применяется, если ковариационные матрицы групп различаются.
Пример: Классификация видов растений по морфологическим признакам.

4. Математическая основа
ДА строит дискриминантные функции (линейные комбинации предикторов), которые максимизируют различия между группами:

$$D = w_1X_1 + w_2X_2 + ... + w_nX_n + c,$$

где:
- (D) — дискриминантная оценка,
- (w_i) — веса признаков,
- (X_i) — предикторы,
- (c) — константа.

Критерий Фишера: Метод находит веса, которые максимизируют отношение:

$$F = \frac{\text{межгрупповая дисперсия}}{\text{внутригрупповая дисперсия}}.$$

5. Пример применения
Задача: Предсказать, купит ли клиент товар (Да/Нет) на основе возраста, дохода и частоты посещений сайта.
- Шаги:
  1. Построить дискриминантные функции для групп "Да" и "Нет".
  2. Для нового клиента вычислить (D) и отнести его к группе с ближайшим центроидом.

6. Преимущества и ограничения

Преимущества	Ограничения
Простота интерпретации результатов.	Требует нормальности распределения.
Эффективен для малых выборок.	Чувствителен к выбросам.
Работает с категориальными и количественными предикторами.	Предполагает линейную разделимость групп.

7. Сравнение с другими методами
- Логистическая регрессия: Лучше для бинарной классификации, но не даёт информации о вкладе переменных в разделение групп.
- Метод опорных векторов (SVM): Эффективен для нелинейных границ, но сложнее интерпретировать.

8. Программные реализации
- R: `lda()` (пакет `MASS`).
- Python: `LinearDiscriminantAnalysis` (библиотека `scikit-learn`).

Перед применением ДА проверяем:
- Нормальность распределения предикторов.
- Однородность дисперсий (для LDA).
- Отсутствие мультиколлинеарности.

5.2 **Ключи к тесту**

1. d
2. a
3. b
4. c
5. a
6. b
7. d
8. b
9. b
10. c
11. b
12. a
13. c
14. a
15. d
16. c
17. b, c, d
18. a