

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Мурманский арктический государственный университет»
(ФГБОУ ВО «МАГУ»)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1.В.ДВ.06.02 Анализ данных

(название дисциплины (модуля) в соответствии с учебным планом)

**основной профессиональной образовательной программы
по направлению подготовки**

**01.03.02 Прикладная математика и информатика
направленность (профиль) Системное программирование и компьютерные технологии**

(код и наименование направления подготовки
с указанием направленности (наименования магистерской программы))

высшее образование – бакалавриат

уровень профессионального образования: высшее образование – бакалавриат / высшее образование – специалитет,
магистратура / высшее образование – подготовка кадров высшей квалификации

бакалавр

квалификация

очная

форма обучения

2023

год набора

Составитель(и):

Ляш Олег Иванович,
доцент, канд. пед. наук,
зав. кафедрой математики, физики
и информационных технологий

Утверждено на заседании кафедры
математики, физики и информационных
технологий факультета
математических и естественных наук
(протокол № 07 от 02.03.2023)

Зав. кафедрой  Ляш О.И.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ) сформировать у студентов представление о наборе технологий и методов интеллектуального анализа (Data Mining), предназначенных для обработки больших объемов исходных данных за счет автоматизации процесса извлечения новых, корректных и потенциально полезных знаний.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

В результате освоения дисциплины (модуля) формируются следующие компетенции:

Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
<p>УК-1: Способен осуществлять поиск, критический анализ информации, применять системный подход для решения поставленных задач</p>	<p>УК-1.1 Рассматривает возможные варианты решения задачи, оценивая их достоинства и недостатки.</p> <p>УК-1.2 Анализирует задачу, выделяя ее базовые составляющие, осуществляет декомпозицию задачи.</p> <p>УК-1.3 Выявляет естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности и способен привлечь для их решения соответствующий математический аппарат</p>	<p><i>Знать:</i></p> <ul style="list-style-type: none"> – принципы мониторинга как информационной технологии получения данных; – типы шкал, в которых могут быть представлены данные, и ограничения на задачи, которые могут быть решены с учётом этих шкал; – основные способы графического и табличного представления данных; – критерии определения аномальных значений в выборочных данных; – основные понятия корреляционного и регрессионного анализа; – причины возникновения ложной корреляции и способы её выявления; основные правила проверки значимости и интервального оценивания уравнения и коэффициентов регрессии; – основные методы анализа и прогнозирования временных рядов; – основные понятия Data Science (науки о данных); – понятия и методы кластерного и классификационного анализа; – основные виды расстояний между объектами (метрик пространства) и кластерами, их особенности и возможности применения; – эквивалентные формулировки задачи о построении базиса признакового пространства из главных компонент и задачи о снижении размерности пространства; – особенности методов интеллектуального анализа текстовых данных. <p><i>Уметь:</i></p> <ul style="list-style-type: none"> – выбирать средства анализа, наиболее эффективные для конкретных данных с учётом их природы, погрешности, пространственного и временного разрешения, а также задач исследования; – правильно понимать и интерпретировать полученные результаты исследования; – уметь критически оценивать возможности и ограничения используемых методов; – применять методы первичной обработки данных; – осуществлять дискретизацию непрерывных данных с учётом решаемой задачи; – выбирать наиболее подходящий способ табличного или графического представления данных, исходя из целей исследования; – проверять наличие статистически значимой линейной связи между переменными; использовать модели временных рядов, выполнять их параметрическую идентификацию, оценивать качество аппроксимации реальных данных выбранной моделью;

		<ul style="list-style-type: none"> – выполнять декомпозицию временных рядов в рамках аддитивной модели; выделять гармонические и квазигармонические аддитивные компоненты временных рядов с помощью Фурье-анализа, вейвлет-анализа, сингулярного спектрального анализа и декомпозиции Хуанга; – выполнять анализ и прогнозирование временных рядов с помощью статистических моделей, сингулярного спектрального анализа и нейронных сетей; выбирать метод классификации или кластеризации в зависимости от цели исследования и характера имеющихся данных; проводить сравнительный анализ различных способов классификации и кластеризации множества объектов с использованием функционалов качества его разбиения; – использовать сингулярный анализ для получения главных компонент и оценки объяснённой дисперсии каждой компоненты; – интерпретировать главные компоненты в терминах предметной области. <p><i>Владеть:</i></p> <ul style="list-style-type: none"> – технологиями поиска данных и оценки их качества; – навыками сбора первичной информации, организации и хранения данных для конкретного исследования; – навыками самостоятельного проведения исследований; – основными категориями, понятиями, методами современной описательной статистики, методами корреляционного анализа; – технологией получения уравнения регрессии и проверки его статистической значимости; – технологией работы с программными средствами, позволяющими осуществлять статистический анализ; – основными приёмами и методами классификации, кластерного и дискриминантного анализа в зависимости от характера используемой информации; – навыками идентификации и интерпретации связи признаков переменных и главных компонент; – технологией работы с программными средствами, позволяющими осуществлять декомпозицию и прогнозирование временных рядов; – технологией работы с программными средствами, позволяющими осуществлять кластерный, классификационный, факторный и дискриминантный анализ; – навыками работы с библиотеками, предоставляющими доступ к нейросетевым моделям (TensorFlow, Keras и др.).
--	--	---

3. УКАЗАНИЕ МЕСТА ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОСНОВНОЙ ПРОФЕССИОНАЛЬНОЙ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина (модуль) «Анализ данных» относится к части, формируемой участниками образовательных отношений, образовательной программы по направлению подготовки 01.03.02 Прикладная математика и информатика направленность (профиль) Системное программирование и компьютерные технологии.

4. ОБЪЕМ ДИСЦИПЛИНЫ (МОДУЛЯ) В ЗАЧЕТНЫХ ЕДИНИЦАХ С УКАЗАНИЕМ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ЧАСОВ, ВЫДЕЛЕННЫХ НА КОНТАКТНУЮ РАБОТУ ОБУЧАЮЩИХСЯ С ПРЕПОДАВАТЕЛЕМ (ПО ВИДАМ УЧЕБНЫХ ЗАНЯТИЙ) И НА САМОСТОЯТЕЛЬНУЮ РАБОТУ ОБУЧАЮЩИХСЯ

Общая трудоемкость дисциплины (модуля) составляет 3 зачетные единицы или 108 часов (из расчета 1 ЗЕ = 36 часов).

Курс	Семестр	Трудоёмкость в ЗЕ	Общая трудоёмкость (часов)	Контактная работа			Всего контактных часов	Из них в интерактивной форме	Кол-во часов на СРС		Кол-во часов на контроль	Форма контроля
				ЛК	ПР	ЛБ			Общее количество часов на СРС	Из них – на курсовую		
4	7	3	108	18	34		52	8	56			зачет
Итого		3	108	18	34		52	8	56			зачет

Интерактивная форма реализуется в виде кейс-заданий по тематикам дисциплины.

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ), СТРУКТУРИРОВАННОЕ ПО ТЕМАМ (РАЗДЕЛАМ) С УКАЗАНИЕМ ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ЧАСОВ И ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ

№ п/п	Наименование раздела, темы	Контактная работа			Всего контактных часов	Из них в интерактивной форме	Кол-во часов на СРС	Кол-во часов на контроль
		ЛК	ПР	ЛБ				
1.	Основные понятия технологий сбора и систематизации данных	1	–	–	1	1	6	–
2.	Предобработка данных, визуализация, первичный статистический анализ	–	–	4	4	1	6	–
3.	Корреляционный и регрессионный анализы	1	–	4	5	1	6	–
4.	Классификация	3	–	4	7	1	6	–
5.	Кластерный анализ	3	–	4	7	1	6	–
6.	Метод главных компонент	2	–	4	6	1	6	–
7.	Анализ и прогнозирование временных рядов	4	–	6	10	1	10	–
8.	Нейросетевое моделирование и глубокое обучение	4	–	8	12	1	10	–
	Зачет							–
	Итого:	18	–	34	52	8	56	–

Содержание дисциплины (модуля)

Раздел 1. Основные понятия технологий сбора и систематизации данных.

Данные, информация, знания. Цели и задачи анализа данных. Источники данных, мониторинг как информационная технология. Суть многомерности данных. Линейное пространство признаков. Пространственно-временные данные. Основные понятия Data Science.

Раздел 2. Предобработка данных, визуализация, первичный статистический анализ.

Детерминированные и случайные составляющие в данных. Анализ качества данных. Дискретизация, удаление выбросов, заполнение пропусков, фильтрация, сглаживание данных. Выборка и генеральная совокупность. Расчёт описательных статистик выборки, анализ закона распределения одномерных случайных величин.

Раздел 3. Корреляционный и регрессионный анализы.

Цель корреляционного анализа, выборочный парный коэффициент корреляции Пирсона (статистическая значимость, согласованность для генеральной совокупности, интервальная оценка). Коэффициенты корреляции Спирмена, Кенделла, Метьюса. Таблица сопряжённости. Цель регрессионного анализа. Метод наименьших квадратов. Получение уравнения одномерной линейной регрессии. Проверка значимости уравнения линейной регрессии.

Раздел 4. Классификация.

Обучение с учителем. Размеченная выборка. Понятие эталона и особенностей в данных. Деревья решений, случайный лес, метод ближайших соседей. Логистическая регрессия. Наивный байесовский классификатор. Метрики качества классификации.

Раздел 5. Кластерный анализ.

Обучение без учителя. Кластерный анализ многомерной выборки: основные понятия, метрики, функционалы и критерии качества, критерии останова. Метод К-средних, метод Expectation-Maximization, плотностные методы, метод кластеризации на базе карт (нейронных сетей) Кохонена.

Раздел 6. Метод главных компонент.

Задача снижения размерности данных. Стандартизация матрицы «объект-признак». Сингулярный анализ. Анализ главных компонент и факторный анализ.

Раздел 7. Анализ и прогнозирование временных рядов.

Свойства временных рядов: аксиоматические и проверяемые. Задачи исследования временных рядов. Декомпозиция одномерного временного ряда на аддитивные составляющие: Фурье-анализ и вейвлет-анализ. Аддитивная модель временного ряда. Декомпозиция временного ряда (одномерный и многомерный случаи) на аддитивные составляющие: сингулярный спектральный анализ и декомпозиция на эмпирические моды Хуанга. Прогнозирование временных рядов по результатам сингулярного спектрального анализа (метод «Гусеница»). Метод Хольта-Уинтерса. Стационарность случайного процесса и порождённого им временного ряда. Статистические модели одномерных временных рядов: AR, MA, ARIMA, ARX, TARX, GARCH.

Раздел 8. Нейросетевое моделирование и глубокое обучение.

Нейронные сети прямого распространения: модель МакКаллока-Питтса, перцептрон Розенблатта, многослойный перцептрон, функция активации, инициализация Нгуен-Видроу, правило обучения Хебба. Моделирование логических функций AND, OR, XOR с помощью нейронов МакКаллока-Питтса. Однослойный и многослойный перцептроны как классификаторы. Многослойный перцептрон с непрерывной функцией активации. Примеры функций активации. Обучение многослойного перцептрона методом обратного распространения ошибки (вывод формул для двухслойного перцептрона). Обучение многослойного перцептрона: метод Левенберга-Марквардта. Прогнозирование одномерного временного ряда с использованием двухслойного перцептрона: план решения задачи. Рекуррентные и свёрточные нейронные сети. Нейросетевые абстракции в компьютерном зрении, анализе текстов, распознавании речи.

6. ПЕРЕЧЕНЬ УЧЕБНО-МЕТОДИЧЕСКОГО ОБЕСПЕЧЕНИЯ, НЕОБХОДИМОГО ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

Основная литература:

1. Анализ данных. Учебник для академического бакалавриата/ В.С. Мхитарян - Отв. ред.- М.:Издательство Юрайт: Национальный исследовательский университет "Высшая школа экономики" (НИУ ВШЭ) (г. Москва), 2016. – 490 с.; То же [Электронный ресурс]. – URL: <http://www.ura.it.ru/book/AF1D197F-1759-422E-9593-8B43E2D1093B>
2. Большаков, А. А. Методы обработки многомерных данных и временных рядов : Учебн. пособие для вузов / А. А. Большаков, Р. Н. Каримов. – 2-е изд., стереотип.. – М. : Горячая линия – Телеком, 2016. – 522 с.
3. Воронцов, К.В. Математические методы обучения по прецедентам [Электронный ресурс] / К.В. Воронцов. – 2011. – URL: <http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf>
4. Подкорытова, О. А. Анализ временных рядов : учебное пособие для бакалавриата и магистратуры / О. А. Подкорытова, М. В. Соколов. – М. : Издательство Юрайт, 2017. – 266 с. – (Бакалавр и магистр. Модуль.); То же [Электронный ресурс]. – URL: <https://www.ura.it.ru/book/55D42DD2-6388-4D0E-87D0-4388738A42BB>
5. Яцков Н. Н. Интеллектуальный анализ данных : учеб. пособие для студ. вузов, обуч. по спец. "Радиофизика", "Прикл. информатика", "Компьютерная безопасность" / Яцков Н. Н. ; Белорусский гос. ун-т. – Минск : БГУ, 2014. – 151 с.

Дополнительная литература:

6. Голяндина, Н. Э. Метод «Гусеница»-SSA : анализ временных рядов : Учеб. Пособие [Электронный ресурс] / Н. Э. Голяндина. – СПб. : С.-Петербургский государственный университет, 2004. – 76 с.
7. Голяндина, Н. Э. Метод «Гусеница»-SSA : прогноз временных рядов : Учеб. Пособие [Электронный ресурс] / Н. Э. Голяндина. – СПб. : С.-Петербургский государственный университет, 2004. – 52 с.
8. Мхитарян, В.С. Анализ данных : Учебник для академического бакалавриата/ В.С. Мхитарян - Отв. ред.-М.: Издательство Юрайт: Национальный исследовательский университет "Высшая школа

экономики" (НИУ ВШЭ) (г. Москва), 2016. – 490 с.; То же [Электронный ресурс]. – URL: <http://www.uraif.ru/book/AF1D197F-1759-422E-9593-8B43E2D1093B>

9. Низаметдинов, Ш.У. Анализ данных : учебное пособие [Электронный ресурс] / Ш.У. Низаметдинов, В.П. Румянцев. – М. : МИФИ, 2013. – 286 с
10. Орлов, Ю. Н. Нестационарные временные ряды : Методы прогнозирования с примерами анализа финансовых и сырьевых рынков / Ю. Н. Орлов, К. П. Осминин. – М. : Либроком, 2011. – 384 с.

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ).

В образовательном процессе используются:

- учебные аудитории для проведения учебных занятий, оснащенные оборудованием и/или его виртуальными аналогами (если используются, если нет - убрать) и техническими средствами обучения: учебная мебель, ПК, оборудование для демонстрации презентаций, наглядные пособия;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду МАГУ.

7.1 ПЕРЕЧЕНЬ ЛИЦЕНЗИОННОГО И СВОБОДНО РАСПРОСТРАНЯЕМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ:

7.1.1. Лицензионное программное обеспечение отечественного производства: нет

7.1.2. Лицензионное программное обеспечение зарубежного производства:

- Mathematica
- MathType
- MS Office
- Statistica

7.1.3. Свободно распространяемое программное обеспечение отечественного производства:

DJVuReader

7.1.4. Свободно распространяемое программное обеспечение зарубежного производства:

Adobe Reader

Mozilla FireFox

7.2 ЭЛЕКТРОННО-БИБЛИОТЕЧНЫЕ СИСТЕМЫ:

- ЭБС «Издательство Лань» [Электронный ресурс]: электронная библиотечная система / ООО «Издательство Лань». – Режим доступа: <https://e.lanbook.com/>;
- ЭБС «Электронная библиотечная система ЮРАЙТ» [Электронный ресурс]: электронная библиотечная система / ООО «Электронное издательство ЮРАЙТ». – Режим доступа: <https://uraif.ru/>;
- ЭБС «Университетская библиотека онлайн» [Электронный ресурс]: электронно-периодическое издание; программный комплекс для организации онлайн-доступа к лицензионным материалам / ООО «НексМедиа». – Режим доступа: <https://biblioclub.ru/>.

7.3 СОВРЕМЕННЫЕ ПРОФЕССИОНАЛЬНЫЕ БАЗЫ ДАННЫХ:

- Информационно-аналитическая система SCIENCE INDEX
- Электронная база данных Scopus
- Базы данных компании CLARIVATE ANALYTICS

7.4. ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ:

- Справочно-правовая информационная система Консультант Плюс <http://www.consultant.ru/>
- ООО «Современные медиа технологии в образовании и культуре» <http://www.informio.ru/>

8. ИНЫЕ СВЕДЕНИЯ И МАТЕРИАЛЫ НА УСМОТРЕНИЕ ВЕДУЩЕЙ КАФЕДРЫ.

Не предусмотрено.

9. ОБЕСПЕЧЕНИЕ ОБРАЗОВАНИЯ ДЛЯ ЛИЦ С ОВЗ.

Для обеспечения образования инвалидов и лиц с ограниченными возможностями здоровья реализация дисциплины может осуществляться в адаптированном виде, с учетом специфики освоения и дидактических требований, исходя из индивидуальных возможностей и по личному заявлению обучающегося.