Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Механико-математический факультет

УТВЕРЖДЕНО:
Декан
Л. В.Гензе

Рабочая программа дисциплины

Современные методы анализа и визуализации больших данных

по направлению подготовки

01.04.01 Математика

Направленность (профиль) подготовки:
Моделирование и цифровые двойники

Форма обучения
Очная

Квалификация
Магистр

Год приема
2025

СОГЛАСОВАНО:
Руководитель ОП
Е.И. Гурина

Председатель УМК
Е.А. Тарасов

Томск – 2025

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ПК-2 Способен проводить тестирование, валидацию и анализ данных цифровых двойников для обеспечения их корректной работы, оптимизации процессов и принятия решений.

ПК-4 Способен документировать процессы разработки и эксплуатации цифровых двойников, работать в команде и взаимодействовать с заказчиками и специалистами для успешной реализации проектов..

УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК 2.2 Анализирует и интерпретирует данные, полученные от цифровых двойников, для принятия предиктивных решений и оптимизации процессов.

ИПК 4.3 Организует и координирует работу команды для достижения поставленных целей проекта.

ИУК 1.1 Выявляет проблемную ситуацию, на основе системного подхода осуществляет её многофакторный анализ и диагностику.

ИУК 1.2 Осуществляет поиск, отбор и систематизацию информации для определения альтернативных вариантов стратегических решений в проблемной ситуации.

ИУК 1.3 Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствий.

2. Задачи освоения дисциплины

– Познакомить обучающихся с основными концепциями, методами и технологиями работы с большими данными.

– Сформировать у обучающегося начальные навыки анализа данных и навыки по визуализации, репрезентации полученных в ходе анализа результатов.

– Сформировать у обучающегося навыки программирования на языке Python для использования внешних модулей языка в рамках решения поставленных задач по анализу и визуализации данных (в том числе и визуализация решений или результатов, оформление решений в виде отчётного ноутбука и др.).

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплины (модули)».

Дисциплина относится к обязательной части образовательной программы. Дисциплина входит в модуль Данные для Digital Twins.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Второй семестр, зачет с оценкой

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

Для успешного освоении дисциплины требуются знания и навыки начального уровня использования ЯП Python и его основных библиотек: numpy, pandas, matplotlib (или других аналогов для визуализации)

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 4 з.е., 144 часов, из которых:
-лекции: 16 ч.
-практические занятия: 16 ч.
в том числе практическая подготовка: 16 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Анализ

Модуль pandas. Структура данных pandas, чтение и запись, работа со структурами pandas, работа с датами, вычисление статистик, корреляционный анализ и доверительные интервалы.

Модуль scipy. Статистическая кластеризация, работа с пространственными структурами данных, статистика.

Модуль sklearn. Подготовка данных для моделирования: методы предобработки данных, типы данных, разбиение, разведывательный анализ данных.

Тема 2. Визуализация

Работа с библиотекой matplotlib. Основные элементы графика, компоновка нескольких графиков, типы графиков, задание цвета и цветовой схемы, цветовая полоса, анимация. Обзор на другие библиотеки: plotly, seaborn.

Тема 3. Решение задачи регрессии.

Линейная регрессия, МНК, Lasso (L1-норма), Ridge (L2-норма). Оценка качества, метрики. Дилемма смещения/дисперсии. Переобучение/недообучение.

Тема 4. Решение задачи кластеризации.

Задачи и подходы кластеризации. Условия задач кластеризации. Алгоритмы кластеризации: K-Means, EM-алгоритм, Агломеративная кластеризация, DBSCAN. Сравнение алгоритмов.

Оценка качества: внутренние и внешние оценки.

Тема 5. Введение в большие данные

Что такое big data. 3V: объём, разнообразие, скорость. Источники данных. Методы и технологии работы с большими данными. Hadoop, Spark. Параллельные вычисления и обработка данных с dask.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля выполнения индивидуальных заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

При оценке выполнения индивидуальных заданий учитывается правильность, оригинальность и сроки выполнения.

По темам 1, 2, 3 и 4 каждый студент получает индивидуальное задание. Оно включает в себя некий результирующий итог по освоению материала соответствующей

темы курса. Работа оформляется в виде отчёта, который студенту необходимо защитить: рассказать о ходе выполнения работы и ответить на дополнительные вопросы по теории выбранной для решения задачи.

По результатам защиты индивидуальных заданий определяется оценка.

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/

Примерный перечень теоретических вопросов

1. Какие описательные статистики есть, и какую полезную информацию с их помощью можно узнать о данных?
2. Что такое корреляции? Какие бывают корреляции? Какую полезную информацию содержит в себе коэффициент корреляции?
3. Что такое доверительный интервал и что мы с его помощью оцениваем?
4. Что такое разведывательный анализ данных? Перечислите его основные этапы.
5. Какие проблемы в данных могут обнаружиться при их исследовании?
6. Какие способы предобработки данных вам известны? Продемонстрируйте на примере (указать не менее 3 шт.)
7. Задача регрессии - что это? Какие типы регрессий вы знаете? В чем их основное отличие? По каким метрикам можно оценить решение задачи регрессии?
8. Задача кластеризации - что это? В чём различие основных подходов в реализации методов кластеризации? Как мы можем оценить работу метода кластеризации?

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=6765

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

в) Самостоятельная работа студентов включает в себя: теоретическое освоение лекционного курса, практическое выполнение заданий и индивидуальных заданий, подготовку к зачету с оценкой. Для выполнения самостоятельной работы обеспечивается доступ к информационным ресурсам курса:

- материалы лекций;
- список вопросов для самостоятельной проверки знаний и подготовки к зачёту.
- список литературы, включающий учебники и книги по изучаемым в курсе вопросам.

Все лабораторные работы и индивидуальные задания подобраны так, чтобы максимально стимулировать психологическую установку студентов-математиков на формирование связи между математической теорией и ее практическим применением. Отчет по каждой лабораторной работе включает теоретическую часть, выполненное практическое задание и анализ полученных результатов.

г) Для успешного освоения материала студентам необходимо посещать занятия, а во время самостоятельной работы пользоваться основной и дополнительной литературой, базами данных и информационно-справочными системами, которые представлены в списке литературы. Самостоятельная работа студентов состоит в повторении материала с практических занятий и самостоятельного изучения дополнительных вопросов, более глубокого анализа темы с помощью литературы.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Элбон К. Машинное обучение с использованием Python. Сборник рецептов. СПб: БХВ, 2020 – 384с.

– Брюс П. Практическая статистика для специалистов Data Science. 2-е. изд., перераб. и доп. – СПб.: БХВ-Петербург, 2021. – 352 с.

– Груздев А. В. Предварительная подготовка данных в Python: Том 1. Инструменты и валидация. – М.: ДМК Пресс, 2023. – 816 с.

– Груздев А. В. Предварительная подготовка данных в Python. Том 2: План, примеры и метрики качества. – М.: ДМК Пресс, 2023. – 814 с.

– Уилке К. Основы визуализации данных: пособие по эффективной и убедительной подаче информации / Клаус Уилке ; [перевод с английского М.А. Райтмана]. — Москва : Эксмо, 2024. — 352 с.

– Кристиан Хилл. Научное программирование на Python / пер. с анг. А. В. Снастина. – М.: ДМК Пресс, 2021. – 646 с.

– Devpractice Team. Библиотека Matplotlib. - devpractice.ru. 2019. - 100 с.

– Devpractice Team. Pandas. Работа с данными. 2-е изд. - devpractice.ru. 2020. - 170 с.

– Маккинни У. Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter / пер. с англ. А. А. Слинкина. 3-е изд. – М.: ДМК Пресс, 2023. – 536 с.

б) дополнительная литература:

– Дейтел П., Дейтел Х. Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.

– Даббас Э. Интерактивные дашборды и приложения с Plotly и Dash. Используем полноценный веб-фреймворк в Python на всю мощь – без JavaScript / пер. с англ. А. Ю. Гинько. – М.: ДМК Пресс, 2022. – 306 с.

– Расширенная аналитика с PySpark: Пер. с англ. – СПб.: БХВ-Петербугр, 2023. – 224 с.

–Нидхем М., Ходлер Э. Графовые алгоритмы. Практическая реализация на платформах Apache Spark и Neo4j. / пер. с англ. В. С. Яценкова – М.: ДМК Пресс, 2020. – 258 с.

– Коритес Б. Графика на Python / пер. с англ. И. Л. Люско. – М.: ДМК Пресс, 2024. – 378 с.

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– операционная система Windows 7 или Windows 10 https://www.microsoft.com/ru-ru/software-download/windows10

– python (дистрибутив python) https://www.python.org/?downloads

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Для проведения лабораторных работ и самостоятельной работы используются ауди-тории учебно-вычислительной лаборатории ММФ.

При выполнении индивидуальных заданий, самостоятельных и лабораторных работ используется свободное и лицензионное программное обеспечение:

– офисный пакет Microsoft Office 2010 (составление отчетов);

– IDE для python (программа для организации работы на языке python).

15. Информация о разработчиках

Стребкова Екатерина Александровна, ст. преподаватель кафедры вычислительной математики и компьютерного моделирования ММФ ТГУ