МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ

Директор Инженерной школы ядерных технологий

\_\_\_\_\_\_\_\_\_\_\_ О. Ю. Долматов

 «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2021 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

**ПРИЕМ 2021 г.**

**ФОРМА ОБУЧЕНИЯ очная**

|  |
| --- |
| **Управление большими данными**  |
|  |  |
| Направление подготовки/ специальность | 01.03.02Прикладная математика и информатика |
| Образовательная программа (направленность (профиль))  | Прикладная математика в инженерии |
| Специализация | Прикладная математика в инженерии |
| Уровень образования | высшее образование - бакалавриат |
|  |  |
| Курс | IV | семестр | **7** |
| Трудоемкость в кредитах (зачетных единицах) | **3** |
| Виды учебной деятельности | Временной ресурс  |
| Контактная (аудиторная) работа, ч | Лекции | **16** |
| Практические занятия | **0** |
| Лабораторные занятия | **32** |
| ВСЕГО | **48** |
| Самостоятельная работа, ч | **60** |
| ИТОГО, ч | **108** |
|  |  |
|  |  |
| Вид промежуточной аттестации | **Экзамен**  | Обеспечивающее подразделение | **ОЭФ ИЯТШ** |
|  |  |
| Руководитель отделения ЭФ |  | Лидер А.М. |
| Руководитель ООП  |  | Крицкий О.Л. |
| Преподаватель |  | Семенов М.Е.  |

2021 г.

# 1. Цели освоения дисциплины

Целями освоения дисциплины является формирование у обучающихся определенного ООП (п. 5.4 Общей характеристики ООП) состава компетенций для подготовки к профессиональной деятельности.

| **Код компетенции** | **Наименование компетенции** | **Индикаторы достижения компетенций** | **Составляющие результатов освоения (дескрипторы компетенций)** |
| --- | --- | --- | --- |
| **Код индикатора** | **Наименование индикатора достижения** | **Код индикатора** | **Наименование** |
| ОПК(У)-1 | Способен применять фундаментальные знания, полученные вобласти математических и (или) естественных наук, и использовать их в профессиональной деятельности | И.ОПК(У)-1.1 | Применяет математический аппарат исследования функций, линейной алгебры, дифференциального и интегрального исчисления, рядов, дифференциальных уравнений, теории функций комплексного переменного в инженерной деятельности | ОПК(У)-1.1В3 | Владеет математическим аппаратом комплексного и операционного исчисления, дифференциальными уравнениями и рядами для проведения теоретического исследования и моделирования физических и химических процессов и явлений, а также, для решения профессиональных задач |
| ОПК(У)-1.1У3 | Умеет решать обыкновенные дифференциальные уравнения и их системы, применять аппарат гармонического и комплексного анализа при решении стандартных задач |
| ОПК(У)-1.1З3 | Знает основные определения и понятия теории дифференциальных уравнений, рядов, функции комплексного переменного и операционного исчисления |
| И.ОПК(У)-1.2 | Использует фундаментальный математический аппарат для построения вычислительных схем | ОПК(У)-1.2В1 | Владеет математическим аппаратом для проведения теоретического исследования и моделирования естественно-научных процессов и явлений, а также, для решения профессиональных задач |
| ОПК(У)-1.2У1 | Умеет решать обыкновенные дифференциальные уравнения, применять аппарат математического анализа действительного переменного и комплексного анализа при решении стандартных задач |
| ОПК(У)-1.2З1 | Знает основные определения и понятия теории математического анализа, теории функций комплексного переменного и операционного исчисления |
| ПК(У)-6 | Способен в составе научно- исследовательского или производственного коллектива и решать задачи профессиональной деятельности | И.ПК(У)-6.1 | Формулирует задачи в рамках проекта | И.ПК(У)-6.1 | Формулирует задачи в рамках проекта |
| ПК(У)-6.1У1 | Умеет формулировать проблему, исходя из действующих физико-математических задач и моделей задач, имеющихся ресурсов и экономических ограничений |
| ПК(У)-6.1З1 | Знает методы и инструменты формулировки проблем с учетом их экономической значимости |

# 2. Место дисциплины (модуля) в структуре ООП

Дисциплина относится к вариативной части Блока 1 учебного плана образовательной программы.

# 3. Планируемые результаты обучения по дисциплине

После успешного освоения дисциплины будут сформированы результаты обучения:

|  |  |
| --- | --- |
| **Планируемые результаты обучения по дисциплине** | **Компетенция** |
| **Код** | **Наименование** |
| РД-1 | Выполнять исследования процессов создания, накопления и обработки информации, включая анализ и создание моделей данных и знаний, языков их описания и манипулирования. | И.ОПК(У)-1.1 |
| РД-2 | Владеть методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской и профессиональной деятельности. | И.ОПК(У)-1.2 |
| РД-3 | Владение методами и инструментами визуализации и ординации многомерных объектов | И.ПК(У)-6.1 |

# Оценочные мероприятия текущего контроля и промежуточной аттестации представлены в календарном рейтинг-плане дисциплины.

# 4. Структура и содержание дисциплины

**Основные виды учебной деятельности**

|  |  |  |  |
| --- | --- | --- | --- |
| **Разделы дисциплины** | **Формируемый результат обучения по дисциплине** | **Виды учебной деятельности** | **Объем времени, ч.** |
| **Раздел (модуль) 1. Стратегии работы с большими объемами данных** | РД-1 | Лекции | **4** |
| Практические занятия | **0** |
| Лабораторные занятия | **12** |
| Самостоятельная работа | **20** |
| **Раздел (модуль) 2. Обработка больших объемов данных** | РД-2, РД-3 | Лекции | **6** |
| Практические занятия | **0** |
| Лабораторные занятия | **12** |
| Самостоятельная работа | **20** |
| **Раздел (модуль) 3. Визуализация и ординация многомерных объектов** | РД-2, РД-3 | Лекции | **6** |
| Практические занятия | **0** |
| Лабораторные занятия | **8** |
| Самостоятельная работа | **20** |

Содержание разделов дисциплины:

**Раздел 1. *Стратегии работы с большими объемами данных***

Форматы данных и типы моделей. Преобразование форматов. Структурированный язык запросов. Широкий и длинный формат данных. Предпроцессинг. Манипуляция. Чтение/запись, кэширование. Функции чтения/записи данных из/в текстовый файл. Справочная система R/Python. Платформы Hadoop, Spark, библиотеки sparklyr, [modeldb](https://github.com/tidymodels/modeldb), фреймворк h2o. Репрезентативная выборка ограниченного размера. Разбиение данных на части. Выполнение вычислений на стороне базы данных. Кэширование вычислений.

**Темы лекций:**

1. Введение в большие данные. Стратегии работы с большими объемами данных

**Темы лабораторных занятий:**

1. Языки, программные среды, фреймворки для обработки данных.
2. Платформы для выполнения распределенных вычислений.
3. Стратегии работы с большими массивами данных.
4. Кэширование вычислений.

**Раздел 2. *Обработка больших объемов данных***

Числовые и текстовые данные. Отсутствующие данные. Ошибки в данных. Выбросы в данных. Дублирующие наблюдения (строки). Мультиколлинеарность. Цифровизация данных. Методы фильтрации, обертки, смешанные методы. Предсказание значений. График boxplot. Предиктивные модели. Дискретная, непрерывная, категориальная переменная. Шкалы. Машинное обучение с учителем, без учителя. Переобучение. Метод опорных векторов. Метод k-ближайших соседей. Выборочные характеристики положения и их использование для заполнения пропусков. Графический способ определения выбросов. Принципы разделения выборки на обучающую и тестовую. Методы преобразования непрерывных данных в категориальные. Классификация и регрессия. Линейные, нелинейные модели.

**Темы лекций:**

2. Базовые алгоритмы интеллектуального анализа данных (Data mining). Подготовка исходных данных к обработке.

1. Выбор признаков (Feature Selection), экземпляров (Instance Selection), дискретизация для классификации (Discretization).

**Темы лабораторных занятий:**

3. Графический разведочный анализ данных.

1. Статистические методы заполнения пропусков и определения выбросов.
2. Определение важности признаков.
3. Классификация и регрессия. Линейные, нелинейные модели.

***Раздел 3. Визуализация* *и ординация многомерных объектов***

Снижение размерности. Метод главных компонент. Метод главных координат. Неметрическое многомерное шкалирование, анализ соответствий. Функции многомерного факторного анализа. Кластерный анализ. Дендограммы. Оптимальное проецирование. Ординационные диаграммы.

**Темы лекций:**

4. Алгоритмы сжатия информационного пространства

**Темы лабораторных занятий:**

1. Системы визуализации ggplot2, GGVIS. Ординационные диаграммы.
2. Метод главных компонент, координат.
3. Кластерный анализ. Дендограммы.
4. Неметрическое многомерное шкалирование.

# 5. Организация самостоятельной работы студентов

Самостоятельная работа студентов при изучении дисциплины (модуля) предусмотрена в следующих видах и формах:

* Работа с лекционным материалом, поиск и обзор литературы и электронных источников информации по индивидуально заданной проблеме курса;
* Изучение тем, вынесенных на самостоятельную проработку;
* Поиск, анализ, структурирование и презентация информации;
* Перевод текстов с иностранных языков;
* Выполнение домашних заданий;
* Программные расчеты;
* Подготовка к практическим занятиям;
* Подготовка к лабораторным занятиям;
* Подготовка к оценивающим мероприятиям.

# 6. Учебно-методическое и информационное обеспечение дисциплины

**6.1. Учебно-методическое обеспечение**

1. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. Москва: ДМК Пресс, 2014. — 588 с. // Лань: электронно-библиотечная система. — URL: <https://e.lanbook.com/book/58703>
2. Буховец, А. Г. Алгоритмы вычислительной статистики в системе R: учебное пособие / А. Г. Буховец, П. В. Москалев. — 2-е изд., перераб. и доп. — Санкт-Петербург : Лань, 2015. — 160 с. // Лань: электронно-библиотечная система. — URL: <https://e.lanbook.com/book/68459>
3. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани; перевод с английского С. Э. Мастицкого. — Москва : ДМК Пресс, 2017. — 456 с. // Лань: электронно-библиотечная система. — URL: <https://e.lanbook.com/book/93580>
4. Москвитин, А. А.. Данные, информация, знания: методология, теория, технологии: монография [Электронный ресурс] / Москвитин А. А.. — Санкт-Петербург: Лань, 2019. — 236 с. https://e.lanbook.com/book/113937
5. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. – Электронная книга, адрес доступа: <https://github.com/ranalytics/data-mining>

**Дополнительная литература**

1. Olvera-López J., Carrasco-Ochoa J. Martínez-Trinidad J. F. and Kittler J. (2010). A review of instance selection methods. Artif. Intell. Rev. 34. 133-143. <https://mafiadoc.com/a-review-of-instance-selection-methods-soft-computing-and-_5b054f698ead0ed4758b4586.html>
2. X. Wu et al. (2008) Top 10 algorithms in data mining. Knowl. Inf. Syst. 14. 1–37. <http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf>
3. Лесковец, Юре. Анализ больших наборов данных: пер. с англ. / Ю. Лесковец, А. Раджараман, Дж. Ульман. — Москва: ДМК Пресс, 2016. — 498 с.
4. Фрэнкс, Билл. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики: пер. с англ. / Б. Фрэнкс. — Москва: Манн, Иванов и Фербер, 2014. — 340 с.
5. Хименко, Виталий Иванович. Случайные данные: структура и анализ: учебник / В. И. Хименко. — Москва: Техносфера, 2019. — 424 с.
6. Орельен Жерон, Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. Концепции, инструменты и техники для создания интеллектуальных систем. М.: Вильямс. - 2018. - 688 с.
7. Хэдли Уикем, Гарретт Гроулмунд, «Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование данных». М.: Вильямс. - 2018. - 592 с.
8. Черняк Л. Серьезно о технологиях для Больших Данных // Открытые системы. СУБД, - 2014. - № 1. http://www.osp.ru/os/2014/01/13039646/

**6.2. Информационное и программное обеспечение**

Internet-ресурсы (в т.ч. в среде LMS MOODLE и др. образовательные и библиотечные ресурсы):

* Персональная страница Семенова М.Е. [*http://portal.tpu.ru/SHARED/s/SME/work*](http://portal.tpu.ru/SHARED/s/SME/work)
* Курс «Big Data» [*https://www.coursera.org/specializations/big-data*](https://vk.com/away.php?to=https%3A%2F%2Fwww.coursera.org%2Fspecializations%2Fbig-data&cc_key=)
* Анализ данных в Spark-кластере с помощью пакета dplyr [https://r-analytics.blogspot.com/2021/03/spark-dplyr.html](https://r-analytics.blogspot.com/2020/03/spark-dplyr.html)
* Локальный Spark-кластер: устанавливаем, подключаемся, пробуем [https://r-analytics.blogspot.com/2021/02/spark-r-connect.html](https://r-analytics.blogspot.com/2020/02/spark-r-connect.html)
* Spark и sparklyr для работы с большими данными в R [https://r-analytics.blogspot.com/2021/02/spark-intro.html](https://r-analytics.blogspot.com/2020/02/spark-intro.html)
* Конференция по большим данным и искусственному интеллекту [*https://bigdatadays.ru/ru/*](https://vk.com/away.php?to=https%3A%2F%2Fbigdatadays.ru%2Fru%2F&cc_key=)
* Профессиональные базы данных и информационно-справочные системы доступны по ссылке: <https://www.lib.tpu.ru/html/irs-and-pdb>

Лицензионное программное обеспечение (в соответствии с **Перечнем лицензионного программного обеспечения ТПУ)**:

1. Document Foundation LibreOffice

# 7. Особые требования к материально-техническому обеспечению дисциплины

В учебном процессе используется следующее лабораторное оборудование для практических и лабораторных занятий:

|  |  |  |
| --- | --- | --- |
| **№** | **Наименование специальных помещений** | **Наименование оборудования** |
|  | Аудитория для проведения учебных занятий всех типов, курсового проектирования, консультаций, текущего контроля и промежуточной аттестации (компьютерный класс). 634028, Томская область, г. Томск, Ленина проспект, д. 2 427А | Доска аудиторная настенная - 1 шт.;Шкаф для одежды - 1 шт.;Шкаф для документов - 1 шт.;Комплект учебной мебели на 10 посадочных мест;Компьютер - 11 шт.7-Zip; Adobe Flash Player; AkelPad; Document Foundation LibreOffice; Google Chrome; Microsoft Office 2007 Standard Russian Academic; Mozilla Firefox ESR; PTC Mathcad Prime 6 Academic Floating; Tracker Software PDF-XChange Viewer; WinDjView |
|  | Аудитория для проведения учебных занятий всех типов, курсового проектирования, консультаций, текущего контроля и промежуточной аттестации 634028, Томская область, г. Томск, Ленина проспект, д. 2 434 | Доска аудиторная настенная - 2 шт.;Комплект учебной мебели на 70 посадочных мест;Компьютер - 1 шт.; Проектор - 1 шт.7-Zip; Adobe Acrobat Reader DC; Adobe Flash Player; AkelPad; Cisco Webex Meetings; Document Foundation LibreOffice; Google Chrome; Microsoft Office 2007 Standard Russian Academic; Mozilla Firefox ESR; Tracker Software PDF-XChange Viewer; WinDjView; Zoom Zoom |

Рабочая программа составлена на основе Общей характеристики ООП ТПУ по направлению 01.03.02 «Прикладная математики и информатика» профиля «Прикладная математика в инженерии» (приема 2021 г., очная форма обучения).

Разработчик(и):

|  |  |  |
| --- | --- | --- |
| Должность | Подпись | ФИО |
| доцент |  | Семенов М.Е. |

Программа одобрена на заседании отделения экспериментальной физики ИЯТШ (протокол № 5 от 31.08.2021 г.).

Заведующий кафедрой – руководитель отделения (на правах кафедры) экспериментальной физики ИЯТШ:
д. т. н. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/Лидер А. М./

подпись