

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

Институт дистанционного образования

УТВЕРЖДЕНО:
Директор ИДО - проректор по РДО
М.О. Шепель

Рабочая программа дисциплины

Разведывательный анализ данных

по направлению подготовки

09.04.03 Прикладная информатика

Направленность (профиль) подготовки:
«Науки о данных»

Форма обучения
Очная

Квалификация
Магистр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
Д.Д. Даммер

Председатель УМК
С.Б. Велединская канд.филол.наук,
доцент

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 - способность самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;

ОПК-3 - способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ОПК-6 - способность исследовать современные проблемы и методы прикладной информатики и развития информационного общества;

ПК-1 - способность управлять получением, хранением, передачей, обработкой больших данных;

ПК-2 - способность применять и разрабатывать новые методы и технологии исследования больших данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.2 Знает основы определения взаимосвязей, закономерностей, обобщения и абстрагирования фундаментальных моделей, законов и методик для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;

ИОПК-3.1 Знает основы сбора, обработки и анализа научно-технической информации, необходимой для решения профессиональных задач;

ИОПК-3.3 Знает основы формулирования результатов, полученных в ходе решения исследовательских задач, в виде аналитических обзоров с обоснованными выводами и рекомендациями;

ИОПК-6.1 Знает основы методов анализа прикладной области, информационных потребностей, технологий сбора, накопления, обработки, передачи и распространения информации;

ИПК-1.4 Знает основы разработки предложений по развитию и совершенствованию системы получения, хранения, передачи и обработки больших данных;

ИПК-1.5 Знает основы разработки моделей данных, адаптированных к технологиям больших данных;

ИПК-2.2 Знает основы проведения аналитических исследований по тематике информационных технологий и технологий больших данных.

2. Задачи освоения дисциплины

- Научиться применять алгоритмы и методы EDA.
- Освоить аппарат математической статистики в контексте EDA;
- Научиться проектировать признаки: определять и создавать новые признаки из существующих данных, а также применять методы кодирования категориальных признаков и работы с временными данными;
- Научиться формулировать статистические гипотезы и проводить тестирование значимости;
- Научиться использовать параметрические и непараметрические тесты для анализа данных;
- Научиться проектировать и логировать эксперименты;
- Научиться проводить A/B-тестирование;
- Научиться работать с ноутбуками в Kaggle для выполнения EDA.

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 “Обязательные дисциплины”.

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

Первый 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: “Программирование на Python”

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 4 з.е., 144 часов, из которых:

– лекции: 6 ч.;

– практические занятия: 18 ч.;

в том числе практическая подготовка: 18 ч.

Объем самостоятельной работы студента составляет: 118,55 ч.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Введение в разведывательный анализ данных. Алгоритмы и методы EDA

Жизненный цикл ML-модели и роль EDA в машинном обучении. Алгоритм и методы EDA. Знакомство с данными: винные обзоры. Проверка. БОНУС. EDA одной строкой кода

Тема 2. Математическая статистика в контексте EDA

Описательная статистика. Меры центральной тенденции в Python. Корреляция. Типы корреляций. Корреляция Пирсона. Типы корреляций. Ранговые корреляции. Визуализация корреляций. Матрица корреляций. График рассеивания. Парные отношения в наборе данных.

Тема 3. Проектирование признаков

Создание признаков. Создание признаков. Внешние источники данных. Создание признаков. Работа с форматом «дата-время». Кодирование признаков. Методы. Преобразование признаков. Нормализация. Стандартизация. Отбор признаков. Мультиколлинеарность.

Тема 4. Статистические тесты

Понятие статистической гипотезы. Статистическая значимость. Статистические тесты. Введение. Тесты на нормальность. Параметрические тесты. Непараметрические тесты. Популярные статистические тесты в Python.

Тема 5. Статистические тесты

Необходимость в статистических тестах. Понятие статистической гипотезы. Статистическая значимость. Статистические тесты: введение. Проверка на нормальность.

Параметрические тесты. Непараметрические тесты. Статистические тесты для категориальных признаков. Операции с векторами. Практика. Статистические тесты в контексте EDA.

Тема 6. Проектирование экспериментов

Введение в проектирование экспериментов.

Логирование экспериментов. Знакомство с Comet.ml. Планирование эксперимента и логирование изменений в Comet.ml. Практика.

Тема 7. A/B-тестирование

Суть и сферы применения A/B-тестирования. Алгоритм, принципы и параметры A/B-тестирования. Анализ результатов A/B-тестирования: кумулятивные метрики. Анализ результатов A/B-тестирования: статистические тесты. Анализ результатов A/B-тестирования: доверительные интервалы. Закрепление знаний

Тема 8. Знакомство с Kaggle

Основы Kaggle. Работа с ноутбуками в Kaggle. Участие в соревновании.

Тема 9. EDA + Feature Engineering. Соревнование на Kaggle

Разбираемся с данными. Машинное обучение для самых маленьких. Лучший способ добиться успеха — следовать советам. Побеждаем в соревновании.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине осуществляется путем контроля посещаемости, выполнения домашних заданий, учебных проектов, тестирования по пройденным темам 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

Промежуточная аттестация проводится в формате итогового индивидуального проекта. **Зачет с оценкой в первом семестре** проводится в письменной форме. Продолжительность зачета 1 час.

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

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Скиллфэктори» – <https://skillfactory.ru/>

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине (<https://www.tsu.ru/sveden/education/eduop/>).

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Кузовкова, Т. А. Экономическая безопасность бизнеса в цифровой среде : учебное пособие / Т. А. Кузовкова, Т. Ю. Салютин. — Москва : Ай Пи Ар Медиа, 2023. — 128 с. — ISBN 978-5-4497-2278-2. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/132156.html> — Режим доступа: для авторизир. пользователей. - DOI: <https://doi.org/10.23682/132156>

- Майкл, Портер Конкурентная стратегия: методика анализа отраслей и конкурентов / Портер Майкл ; перевод И. Минервин ; под редакцией О. Нижельской. — 6-е изд. — Москва : Альпина Паблишер, 2020. — 456 с. — ISBN 978-5-9614-5752-0. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/93025.html> — Режим доступа: для авторизир. пользователей
- Малаховская, М. В. Конкурентная разведка : учебное пособие / М. В. Малаховская. — Санкт-Петербург : Интермедия, 2017. — 120 с. — ISBN 978-5-4383-0136-3. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/66790.html> — Режим доступа: для авторизир. пользователей
- Смелик, Р. Г. Экономическая безопасность бизнеса : учебник для бакалавров / Р. Г. Смелик, А. В. Боженкова. — Омск : Издательство Омского государственного университета, 2020. — 201 с. — ISBN 978-5-7779-2446-9. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/108148.html> — Режим доступа: для авторизир. пользователей
- Чернопяттов, А. М. Бенчмаркинг : учебное пособие / А. М. Чернопяттов. — Москва : Ай Пи Ар Медиа, 2023. — 152 с. — ISBN 978-5-4497-1859-4. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: <https://www.iprbookshop.ru/127566.html> — Режим доступа: для авторизир. пользователей

б) дополнительная литература:

- Ахим Бююль, Петер Цефель. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. Торгово-издательский дом Dia Soft Москва, Санкт-Петербург, Киев 2002.
- Маркетинговые исследования с помощью SPSS (4-е издание + CD) Практическое руководство. Нэреш К. Малхотра М.: Вильямс, 2006, - 1200 с.
- Практическая бизнес-статистика Эндрю Сигел, 2008 , Вильямс , 1056с.
- Таганов Д.Н. SPSS: Статистический анализ в маркетинговых исследованиях. — СПб.: Питер, 2005. — 192 с.: ил.
- Jeffrey M. Wooldridge - Introductory Econometrics - A Modern Approach, 2008 Publisher: Thomson Learning

в) ресурсы сети Интернет:

- открытые онлайн-курсы

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

- 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/>
- ЭБС Консультант студента — <http://www.studentlibrary.ru/>
- Образовательная платформа Юрайт — <https://urait.ru/>
- ЭБС ZNANIUM.com — <https://znanium.com/>
- ЭБС IPRbooks — <http://www.iprbookshop.ru/>

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.

15. Информация о разработчиках

Салаева А.С., методист Skillfactory