

Аннотация рабочей программы дисциплины (модуля)
Б1.В.ДВ.03.02 Введение в анализ данных

Цель и задачи дисциплины

Цель дисциплины

Целью дисциплины является изучение представления о сборе, обработке и анализе данных, принципов обработки данных, задачи классификации с обучением, поиска ассоциативных правил, элементами кластерного анализа.

Задачи дисциплины

Основными задачами изучения дисциплины являются:

- изучение современных методов анализа данных;
- изучение проблем, возникающих при анализе данных, и путей их решения;
- сформировать навыки анализа данных различной природы.

Формируемые компетенции и индикаторы их достижения по дисциплине

Код компетенции	Содержание компетенции	Код и наименование индикатора достижения компетенции
ПКС-2	Способен проводить формализацию предметной области с целью создания информационной системы	ПКС-2.1 - Знает требования к компьютерному программному обеспечению; виды технической спецификации на программные компоненты и их взаимодействие; методы проектирование компьютерного программного обеспечения ПКС-2.2 – Умеет применять требования к компьютерному программному обеспечению; разрабатывать технические спецификации на программные компоненты и их взаимодействие; применять методы проектирования компьютерного программного обеспечения; ПК-2.3 – Владеет методами разработки требований к компьютерному программному обеспечению, технических спецификаций на программные компоненты, методами проектирования компьютерного программного обеспечения.

Содержание разделов дисциплины

Тема 1. Математический аппарат (refresher). Введение в модуль NumPy. Основы работы с Pandas. Разведывательный анализ данных.

Математический аппарат для анализа данных: векторы, матрицы, функции и производные. Особенности типов данных в NumPy. Работа с векторами и матрицами. Вычисление главных статистических метрик с помощью NumPy (среднее, медиана, мода, дисперсия). Введение в модуль для работы с числовыми данными NumPy (Numerical Python). Особенности типов данных в NumPy. Работа с векторами и матрицами. Введение в модуль для работы с табличным представлением данных Pandas. Преобразование словарей в табличный формат Pandas, загрузка данных из внешних источников. Особенности фильтрации и обращения к данным.

Тема 2. Визуализация данных. Представление результатов исследования

Введение в визуализацию данных. Нюансы визуализации данных и принципы человеческого восприятия. Правила создания хороших визуализаций. Обзор различных видов графиков (гистограмма, бар-чарт, секторная диаграмма, линейные графики, график рассеяния, тепловая карта и т.д.). Особенности разных видов графиков и их использования.

Тема 3. Работа с текстовыми данными. Сбор данных из открытых источников. Предварительная обработка текстовых данных. Текстовый анализ

Введение в анализ текста. Применение в политологии. Особенности подготовки данных. Анализ текста. Латентное размещение Дирихле.

Тема 4. Анализ сетей

Введение в анализ сетей. Основные метрики и параметры сетей. Введение в модуль NetworkX. Подготовка данных для анализа сетей. Примеры визуализации сетей на примере данных из социальной сети.

Тема 5. Введение в машинное обучение. Модуль sklearn. Задачи классификации и линейные модели.

Введение в машинное обучение. Обучение с учителем и без учителя. Проблема переобучения. Регрессионные модели. Метод наименьших квадратов. Логистическая регрессия. Решение задач кластеризации. Меры расстояния. Обзор алгоритмов кластеризации (иерархические алгоритмы, алгоритмы квадратичной ошибки, выделение связных компонент).

Тема 6 Деревья решений. Случайный лес. Ансамбли моделей

Введение в ансамбли моделей. Стэкинг и бэггинг. Случайный лес. Бустинг. Практикум 14. Разбор примеров. Построение ансамблей моделей на наборе данных “Титаник”.