Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Сибирский институт управления – филиал РАНХиГС
Кафедра бизнес-аналитики и статистики

ПРОГРАММА БАКАЛАВРИАТА

«Информационные технологии в государственном и муниципальном управлении»

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ И ВИЗУАЛИЗАЦИИ ДАННЫХ
Б1.В.24.02

краткое наименование дисциплины: нет

по направлению подготовки:
38.03.04. Государственное и муниципальное управление

формы обучения: очная

Год набора – 2023

Новосибирск, 2022

1.Цель освоения дисциплины:

Цель изучения дисциплины является получение минимально необходимого объема теоретических знаний, а также развитие приобретенных ранее знаний, умений и навыков в области базового курса информационных и цифровых технологий, полученные в средних образовательных учреждениях, на первом курсе образовательной программы.

Дисциплина «Современные методы обработки информации и визуализации данных» является общим теоретическим и методологическим основанием для подготовки к процедуре защиты и защите выпускной квалификационной работы.

2.План курса:

Б1.В.24.02 «Современные методы обработки информации и визуализации данных»

Раздел 1. Основы обработки и анализа данных

Тема 1.1. Введение в анализ данных

Предмет и задачи дисциплины. Состав современного программного обеспечения анализа данных. Основные направления и проблемы анализа массовых данных. Понятие Data Mining. Data Mining как часть рынка интеллектуальных технологий. Набор данных и их атрибутов. Измерения. Шкалы измерения признаков. Задача классификации. Задача регрессии. Задача прогнозирования. Задача кластеризации. Методы визуализации.

Процесс Data Mining. Средства извлечения данных: методы и возможности. Начальные этапы: анализ предметной области; постановка задачи, подготовка данных. Построение и использование модели. Стандарты Data Mining. Стандарт CRISP-DM. Стандарт PMML. Стандарты SQL/MM и OLE DB for Data Mining. Другие стандарты, описывающие методологию Data Mining. Категории статистики: совокупность, статистические показатели, признаки. Подготовка исходных массивов данных.

Тема 1.2. Формирование массивов данных

Технологии формирования наборов данных и информационных массивов. Сэмплинг. Парсинг данных. Типы наборов данных. Форматы хранения данных. Метаданные. Открытые данные. Базы данных.

Статистическое наблюдение, социологический опрос. Цифровые техники проведения наблюдения и сбора данных.

Тема 1.3. Описательный анализ данных

Сущность, значение, виды и способы расчета статистических величин (абсолютных, относительных и средних).

Вариация признака в совокупности, значение ее статистического изучения. Ряд распределения как метод отражения вариации признака. Атрибутивные и вариационные ряды. Дискретные и интервальные вариационные ряды. Основные структурные характеристики вариационного ряда: мода, медиана, квартили, децили, перцентили. Основные показатели вариации: размах вариации, среднее линейное отклонение, дисперсия, среднее квадратическое отклонение, коэффициент вариации. Правило сложения дисперсий. Эмпирический коэффициент детерминации. Эмпирическое корреляционное отношение.

Эмпирическое распределение. Теоретические распределения. Показатели эксцесса и асимметрии. Статистические критерии оценки согласованности эмпирических и теоретических распределений.

Тема 1.4. Подготовка и обработка данных. Качество данных

Принципы формирования массивов данных. Этапы подготовки данных. Проблемы и ошибки. Очистка данных. Обработка пропущенных значений (импутация данных). Обработка дубликатов. Группировка данных. Квантование. Виды группировок. Определение числа групп и величины интервалов группировки.

Исследование совокупности на наличие выбросов и аномалий. Трансформация данных. Приведение данных к сопоставимому виду. Задача редукции данных. Основные технологии и методы редукции данных.

Тема 1.5. Визуализация данных

Методы визуализации данных в системах обработки статистических данных. Виды и классификация статистических таблиц. Формулировка заголовка таблицы: принципы и обязательные составляющие. Элементы таблицы и этапы ее построения. Электронные таблицы. Сводные таблицы. OLAP- технологии.

История развития графических методов, современные тенденции в использовании графического метода визуализации. Элементы графика и принципы его построения. Отдельные виды графиков и диаграмм: назначение и особенности построения.

Современные приемы визуализации данных. Дашборды. Инфографика в различных предметных областях.

Тема 1.6. Проверка статистических гипотез

Виды статистических гипотез. Ошибки первого и второго рода. Т-тест Стьюдента. Критические значения. Уровень значимости. Статистические гипотезы о равенстве средних.

Тема 1.7. Исследование взаимосвязей.

Понятие о взаимосвязи явлений и процессов. Задачи статистического изучения взаимосвязей. Виды и формы связей. Технологии исследования взаимосвязей переменных разных типов шкал.

Основы корреляционно-регрессионного анализа. Парная и множественная корреляция. Уравнение регрессии. Оценка параметров уравнения регрессии, интерпретация параметров. Оценка статистической значимости параметров парной линейной регрессии.

Непараметрические методы измерения взаимосвязей. Построение таблиц сопряженности. Определение силы связи номинальных переменных. Оценка значимости связи в таблицах сопряженности. Оценка связи ранговых переменных.

Раздел 2. Основы интеллектуального анализа данных

Тема 2.1. Задача регрессии

Регрессия, метрики качества регрессии. Линейная регрессия. Библиотека sklearn. Пример: линейная регрессия на одном признаке, её геометрия, идея обучения. Идея обучения для многомерной регрессии. Градиентный спуск. Проблемы градиентного спуска. Метрики качества. Интерпретируемость линейной модели. Групповая линейная регрессия.

Тема 2.2. Задача классификации данных

Идея классификации. Логистическая регрессия. Меры расстояния. Алгоритмы кластеризации (иерархические алгоритмы, алгоритмы квадратичной ошибки, выделение связных компонент). Бинарные классификации. Множественные классификации. Технологии проведения дискриминантного анализа. Пробит- и логит-регрессии. Нейросетевые техники. Оценка качества проведения классификации, метрики качества. ROC-кривые. Технологии получения деревьев классификации и регрессии.

Тема 2.3. Задача кластеризации данных

Алгоритмы и программы кластерного анализа признаков приемами оптимизации, подготовка исходных данных, процедура решения, анализ результатов. Агломеративные и дивизивные процедуры. Расстояния в признаковом пространстве. Алгоритмы и программы аналитической многомерной группировки, подготовка исходных данных, варианты задания групп, интерпретация результатов.

Тема 2.4. Основы работы с временными рядами и прогнозирования

Основы работы с временными рядами. Индексирование, выборка, подмножества.

Временные ряды с неуникальными индексами. Диапазоны дат, частоты и сдвиг. Генерация диапазонов дат. Частоты и смещения дат. Сдвиг данных с опережением и с запаздыванием. Графики временных рядов. Скользящие функции: экспоненциально взвешенные функции, бинарные скользящие функции.

Тема 2.5. Обработка текстовых данных

Задачи анализа текстовых данных. Обработка естественного языка. Мешок слов. Стемминг и лемматизация. Классификация текстов. Семантический анализ текстов. Семантическое ядро. Показатели анализа текстов: тошнота, доля уникальных и значимых слов.

Основная литература

1. Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2020. — 490 с. — (Высшее образование). — URL: https://urait.ru/bcode/450166

2. Елисеева И. И., Боченина М. В., Капралова Е. Б., Курышева С. В., Нерадовская Ю. В., Флуд Н. А., Науменко Н. В., Ширина А. Н., Парик И. Ю. Статистика. Практикум [Электронный ресурс]:Учебное пособие Для академического бакалавриата. - Москва: Юрайт, 2019. - 514 с – Режим доступа: https://urait.ru/bcode/425262

3. Мхитарян В. С., Агапова Т. Н., Суринов А. Е., Луппов А. Б., Миронкина Ю. Н. Статистика. В 2 ч. Часть 1 [Электронный ресурс]:Учебник и практикум для вузов. - Москва: Юрайт, 2021. - 249 с – Режим доступа: https://urait.ru/bcode/475170