



Силабус дисципліни

«Прикладний аналіз і обробка даних на мові R»



Викладач: *Калініна Ірина Олександрівна*

д-р техн. наук, доцент, в. о. професора кафедри інтелектуальних інформаційних систем

Очікувані результати навчання

В результаті вивчення дисципліни студент

має знати:

- основні методи прикладного інтелектуального аналізу та машинного навчання для обробки реальних даних;
- можливості мови R та середовища RStudio для статистичної обробки даних;
- можливості мови R та середовища RStudio для графічного та візуального аналізу даних;
- основні операції управління та обробки реальних даних;
- основні методи статистичної обробки даних;
- основні етапи вирішення задач машинного навчання.

має вміти:

- проводити візуальний аналіз реальних даних;
- використовувати можливості програмного середовища RStudio для вирішення завдань регресії;
- використовувати можливості програмного середовища RStudio для вирішення завдань класифікації і кластеризації;
- розробляти процедури статистичного машинного навчання для вирішення різноманітних завдань;
- використовувати сучасні методи інтелектуального аналізу для обробки реальних даних.

Пререквізити

Обсяг дисципліни: 4 кредити ECTS (16 годин лекцій, 32 годин практичних робіт).

Мета: набуття знань та навичок прикладного аналізу і обробки реальних даних в різних задачах машинного навчання.

Оригінальність навчальної дисципліни:

Авторський курс.

Зміст дисципліни

Тема 1. Мова програмування R для прикладного аналізу і обробки реальних даних. Особливості мови R, її застосування в науки о даних. Середовище програмування RStudio. Огляд репозиторіїв даних. Робота зі структурами даних за допомогою мови R. Розвідувальний аналіз даних. Розподіл даних та вибірок.

Тема 2. Створення набору даних. Структури даних (вектори, матриці, масиви даних, таблиці даних, фактори, списки). Особливості даних різних типів. Імпорт даних із різних джерел. Анотація даних. Функції роботи з об'єктами. Маніпулювання даними в R. Робота з масивами даних, Застосування матричних операцій. Перетворення масивів. Робота з датафреймами, перетворення, відбір даних. Інтерпретування відібраних даних.

Тема 3. Візуальний аналіз даних за допомогою мови R. Базові графічні можливості R. Огляд основних типів діаграм, гістограм. Функції plot(), hist(), barplot(), dotchart() та їх параметри. Загальні аргументи графічних функцій. Гістограми, функції ядерної густини та функція cdplot(). Діаграми розмахів. Кругові та стовпчикові діаграми. Категоризовані графіки. Додавання тексту, налаштування параметрів осей та умовних позначень. Інтерпретація візуалізації відібраних даних. Розподіл даних та вибірок.

Тема 4. Основи управління реальними даними на мові R. Сортуння даних. Об'єднання наборів даних. Поділ наборів даних на складові. Числові та текстові дані.

Відповідно до освітньої програми необхідно попередньо оволодіти знаннями з дисциплін загальної, практичної та професійної підготовки бакалавра та магістра з галузі знань F «Інформаційні технології».

Пореквізити

Компетентності, знання та вміння, отримані в рамках вивчення даної дисципліни, можуть бути застосовані при проходженні «Передатестаційна практики» та підготовки кваліфікаційної роботи. А також виконанні наукових робіт та у майбутній професійній діяльності.

Семестровий контроль: іспит

Оцінювання:

За семестр: 70 балів

За залік: 30 балів

Види робіт:

Практичні роботи.

Технічне забезпечення

Проекційне мультимедійне обладнання (проектор, екран, ноутбук/комп'ютер).

Доступ до мережі Internet, точка доступу Wi-Fi.

Програмне забезпечення: мови R, середовище RStudio.

Система електронного навчання Moodle 3.9.

Політика щодо дедлайнів

Роботи, які здаються із порушенням термінів без поважних причин, оцінюються на нижчу оцінку.

Політика щодо академічної доброчесності

Передбачає самостійне виконання лабораторних робіт. Списування під час заліку (в т. ч. із використанням мобільних пристроїв) заборонено. У разі виявлення плагіату або списування роботи не зараховуються.

Розв'язання задачі управління даними. Управління виконанням команд. Повторення та цикли. Виконання за умови. Функції написані користувачем.

Тема 5. Основні методи статистичної обробки даних. Описова статистика, підгонка розподілів та суміжні завдання. Таблиці частот та таблиці спряженості. Кореляція. Статистичні експерименти та перевірка значущості. Статистичні тести. Непараметричні випробування між групових відмінностей. Статистика повторних вибірок та бутстреп-аналіз.

Тема 6. Регресія та передбачення. Методи обробки даних середньої складності. Налагодження регресійних моделей. Діагностика регресійних моделей. Незвичайні спостереження. Вибір «найкращої» регресійної моделі. Введення у дисперсійний аналіз.

Тема 7. Класифікація і кластеризація. Стратегії щодо незбалансованих даних. Аналіз головних компонент. Факторний аналіз даних. Наївний алгоритм Байєса. Дискримінантний аналіз. Логістична регресія. Підгонка моделі Оцінювання моделей класифікації. Шалювання та категоріальні змінні.

Тема 8. Статистичне машинне навчання. K найближчих сусідів. Деревоподібні моделі. Гомогенні та гетерогенні ансамблі моделей.

Тема 9. Просунуті графічні методи аналізу даних в R. Пакет lattice. Пакет ggplot2. Пакет playwith. Пакет laticist. Створення інтерактивної графіки за допомогою пакету iplots. Пакет rggobi.

Критерії оцінювання лабораторних/практичних/індивідуальних/робіт/ доповідей/проєктів

Максимальна кількість балів – студент з високою якістю самостійно виконав весь обсяг робіт, відповідає на всі питання, пов'язані з виконаними роботами, та робить додаткові розрахунки, які йому пропонує викладач. У викладача немає претензій щодо реалізації та вимог до виконання роботи.

70%-99% від максимальної кількості балів – студент з достатньою якістю виконав всі завдання, але в процесі роботи він робив деякі помилки, які, після вказування на них викладачем, самостійно виправляв. На деякі питання він відповідає з похибкою. Запропоновані викладачем додаткові розрахунки робить з деякою потугою. Не всі вимоги до виконання роботи дотримані.

40%-69% від максимальної кількості балів – студент самостійно виконав всі роботи, але якість реалізації недостатня (помилки при розрахунках, не всі вимоги до

роботи дотримані). На питання щодо виконання робіт відповідає не зовсім чітко. Є помилки при відповідях.

1%-39% від максимальної кількості балів – студент самостійно виконав не всі роботи, при цьому якість реалізації недостатня (помилки при розрахунках, не дотримується вимог до оформлення роботи). На питання щодо виконання робіт відповідає не чітко. Є грубі помилки при відповідях.

0 балів – студент не виконав весь обсяг робіт, або виконав з грубими помилками. Він має проблеми з розрахунками, не знає теоретичного матеріалу, програмна реалізація не відповідає поставленим вимогам.

Для зарахування практичної роботи здобувач має змогу надати сертифікат (або інший документ) про проходження курсів на платформах *Prometheus, Coursera* тощо та за умови співпадіння направленості курсів блоку освітньої компоненти.