



Силабус курсу ПРИКЛАДНИЙ АНАЛІЗ ДАНИХ

Освітньо-професійна програма «Системний аналіз»
Ступінь вищої освіти – перший (бакалаврський)
Спеціальність: 124 «Системний аналіз»

Рік навчання: 3, Семестр: 5

Кількість кредитів: 5 Мова викладання: українська

Керівник курсу

ПШ

д.т.н., доцент, Роман Пасічник

Контактна інформація roman.pasichnyk@gmail.com

Опис дисципліни

Прикладний аналіз даних служить базою прийняття управлінських рішень. На сучасному етапі його значимість значно зросла у зв'язку із величезним обсягом інформації, який породжує Веб. Такі обсяги інформації важко обробляти без спеціалізованих пакетів, зокрема найпопулярніших Pandas та R, що передбачають володіння відповідними програмними засобами та навичками використання спеціалізованих модулів та бібліотек.

Перш ніж застосовувати методи аналізу, дані повинні бути підготовані та попередньо проаналізовані. Такий аналіз включає перш за все виділення груп однорідних об'єктів методами кластеризації. На наступному етапі відбувається аналіз густин розподілів у виділених однорідних сукупностях. На подальшому етапі з'являється можливість перевірки статистичних гіпотез щодо подібності розподілів та їх характеристик як за параметричними, так і за непараметричними критеріями. І нарешті на заключних етапах з'являється можливість побудови прогнозів поведінки характеристик досліджуваних об'єктів в залежності від прогнозованих оцінок вхідних параметрів. Ці прогнози можуть бути коротко, середньо та довготривалими. Аналіз та застосування згаданих підходів і складає основний зміст даного курсу.

Структура курсу

Години (лек. / модуль.)	Тема	Результати навчання	Завдання
3 / 1	1. Прикладні пакети аналізу даних	Розуміти особливості основних завдань та інструментів аналізу даних, пакету статистичного аналізу Pandas, формування структур та вибір елементів із них, методів сортування словників по ключах та значеннях, пакету R та середовища аналітика RStudio.	Тести, питання
3 / 1	2. Загальний аналіз вибірок із пакетом R	Вміти будувати коробкові діаграми, оцінювати обсяги груп за категоріями, середнє значення, медіана та усічене середнє, характеристики розсіювання, довірчі інтервали для прогнозів, дисперсію, середньоквадратичне розсіювання та міжквартильний розмах.	Тести, завдання
3/1	3. Задача класифікації підвибірок	Вміти додавати ознаки для класифікації, стовпчикові діаграми, реалізовувати задачі ієрархічного кластерного аналізу, метод к-середніх, методи агрегації даних.	Тести, завдання

3/1	4. Ієрархічний кластерний аналіз	Аналізувати віддаль між кластерами, центроїдний метод та його недоліки, методи далекого та найближчого сусіда, методи попереднього кластерного аналізу, метод покрокового агрегування кластерів, дендрограми та алгоритми їх побудови, насипи (лікті) для масштабних кластеризацій.	Тести, завдання
3/1	5. Метод к-середніх	Вміти задавати центри кластерів, кількість проведення процедур кластеризації, вибрати найкращу кластеризацію за критерієм якості, реалізувати метод голосування кількості кластерів.	Завдання
3/2	6. Основні статистичні гіпотези	Перевіряти гіпотезу погодження, гіпотезу нормальності, гіпотезу експоненціальності для часу безвідмовної роботи, гіпотезу однорідності про рівність розподілів, гіпотезу незалежності, гіпотезу про параметр розподілу.	Тести, завдання
3/2	7. Порівняння характеристик вибірок	Реалізувати відсів викидів, перевіряти гіпотези рівності дисперсій, використовувати критерій Флінгена, критерій рівності математичних сподівань Стюдента, аналізувати випадок парних вибірок, реалізувати перевірку однорідності вибірок, перевірку корельованості вибірок, критерій хі квадрат, тести порівняння відносних частот у вибірок, критерій Манна-Уїтні порівняння медіан.	Контрольна робота
3/2	8. Лінійна регресія	Реалізувати перевірку гіпотези про лінійну залежність вибірок, підбір невідомих параметрів, аналіз залишків, пошук колінеарних залежностей, методи заповнення пропусків у даних, усунення змінних, що породжують колінеарність, перевірку гіпотези про рівність нулю коефіцієнта при неінформативних змінних, покроковий метод зменшення моделі.	Тести, завдання
3/2	9. Прогноз на основі лінійної регресійної моделі.	Здійснювати аналіз наявності тренду та сезонності, аналіз викидів, індикатори періодів сезонності, навчання моделі, побудову прогнозу, підбір структури моделі, метод валідації, аналіз коротких сезонних часових рядів.	Тести, завдання
3/2	10. Аналіз часових рядів	Реалізувати перевірку стаціонарності рядів, дослідження автокореляційної та частково автокореляційної функцій, підбір параметрів моделі ARIMA, аналіз залишків моделі, підбір параметрів моделі на основі критерію Акаїки.	Тести, завдання

Літературні джерела

1. Wes McKinney. Python for Data Analysis. // <https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Analysis.pdf>
2. Йорн Гіз. Підручник із ієрархічної кластеризації та дендрограм SciPy. <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/>
3. N D Lewis. Deep Time Series Forecasting with Python: An Intuitive Introduction to Deep Learning for Applied Time Series Modeling. 2016 // <https://www.pdfdrive.com/deep-time-series-forecasting-with-python-an-intuitive-introduction-to-deep-learning-for-applied-time-series-modeling-e187585770.html>.
4. 8host.com. Прогнозування часових рядів за допомогою ARIMA в PYTHON 3. <https://www.8host.com/blog/prognozirovanie-vremennyh-ryadov-s-pomoshhyu-arima-v-python-3/>
5. Joel Grus. Data Science from Scratch. 2019 // https://covid19.uthm.edu.my/wp-content/uploads/2020/04/Data-Science-from-Scratch-First-Principles-with-Python-by-Joel-Grus-z-lib.org_.epub_.pdf.
6. S. Muthukrishnan. Data Streams: Algorithms and Applications. // <https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/Muthu-Survey.pdf>
7. Benjamin Bengfort & Tony Ojeda & Rebecca Bilbro. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning // <https://www.pdfdrive.com/applied-text-analysis-with-python-enabling-language-aware-data-products-with-machine-learning-e176357161.html>

Політика оцінювання

- **Політика щодо дедлайнів та перескладання:** Роботи, які здаються із порушенням термінів безповажних причин, оцінюються на нижчу оцінку (-20 балів). Перескладання модулів відбувається із дозволу деканату за наявності поважних причин (наприклад, лікарняний).
- **Політика щодо академічної доброчесності:** Усі письмові роботи перевіряються на наявність плагіату і допускаються до захисту із коректними текстовими запозиченнями не більше 20%. Списування під час контрольних робіт та екзаменів заборонені (в т.ч. із використанням мобільних девайсів).
- **Політика щодо відвідування:** Відвідування занять є обов'язковим компонентом оцінювання, за яке нараховуються бали. За об'єктивних причин (наприклад, хвороба, міжнародне стажування) навчання може відбуватись в он-лайн формі за погодженням із керівником курсу.

Оцінювання

Остаточна оцінка за курс розраховується наступним чином:

Види оцінювання	% від остаточної оцінки
Модуль 1 (теми 1-5) 1. Усне опитування під час заняття (5 тем по 10 балів = 50 балів) 2. Письмова робота = 50 балів	20
Модуль 2 (теми 6-10) 1. Усне опитування під час заняття (5 тем по 6 балів = 30 балів) 2. Письмова робота = 70 балів	20
КПЗ 1. Написання та захист КПЗ = 80 балів. 2. Виконання завдань під час тренінгу = 20 балів	20
Екзамен – тести, завдання	40

Шкала оцінювання:

За шкалою ЗУНУ	За національною шкалою	За шкалою ECTS
90–100	відмінно	A (відмінно)
85–89	добре	B (дуже добре)
75-84		C (добре)
65-74	задовільно	D (задовільно)
60-64		E (достатньо)
35-59	незадовільно	FX (незадовільно з можливістю повторного складання)
1-34		F (незадовільно з обов'язковим повторним курсом)