

Назва курсу	CSB015: Аналіз даних
Викладач	Пилипенко А.І., к.т.н., доцент
Профайл викладача	https://www.krok.edu.ua/ua/pro-krok/spivrobitniki/pilipenko-anna-ivanivna
Контактний телефон	(+38 050) 195-86-67
E-mail	annapi@krok.edu.ua
Консультації	Згідно з розкладом (online), MS Teams

1. Коротка анотація до курсу

Метою курсу є формування знань та навичок роботи з даними (збір, обробка, візуалізація); постановки дослідницького завдання і тестування гіпотез за допомогою кількісних методів; презентації отриманих результатів.

У курсі для вирішення аналітичних задач використовується мова програмування Python, бібліотеки pandas, numpy, scipy, sklearn, matplotlib. Практичні завдання виконуються в інтерактивному веб-середовищі Jupyter Notebooks.

2. Результати навчання

Програмні компетентності та результати навчання, компоненти яких формуються та розвиваються у курсі

Внесок дисципліни “Аналіз даних” в формування наступних програмних результатів навчання:

Програмний результат навчання	Результат навчання з дисципліни
ПР3 Використовувати знання закономірностей випадкових явищ, їх властивостей та операцій над ними, моделей випадкових процесів та сучасних програмних середовищ для розв’язування задач статистичної обробки даних і побудови прогнозних моделей; (спільно з дисципліною CSB012 Теорія ймовірностей та математична статистика) ПР4 Використовувати методи обчислювального інтелекту, машинного навчання, нейромережевої та нечіткої обробки даних, генетичного та еволюційного програмування для розв’язання задач розпізнавання, прогнозування, класифікації, ідентифікації об’єктів керування тощо. (спільно з дисципліною CSB021 Штучний інтелект) ПР12 Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використанням програмних інструментів підтримки багатовимірного аналізу даних на основі технологій DataMining, TextMining, WebMining. (спільно з дисципліною CSB021 Штучний інтелект)	Студент може: • демонструвати вміння використовувати мову Python для вирішення аналітичних задач; • завантажувати дані і працювати з ними (фільтрація, агрегація, заповнення пропущених значень); • підраховувати описові статистики, оцінювати розподіл, інтерпретувати кореляції; • вибирати коректні графіки для візуалізації даних, вміти кастомізувати їх зовнішній вигляд, інтерпретувати графіки; • будувати прогнозні моделі для різних типів даних (Time series включно); • самостійно опановувати нові спеціалізовані бібліотеки для обробки, візуалізації і аналізу даних.

Передумовою формування результатів навчання з дисципліни є володіння здобувачем на достатньому рівні наступних компетентностей:

Загальні компетентності:

- ЗК3 Знання та розуміння предметної області та розуміння професійної діяльності;
- ЗК7 Здатність до пошуку, оброблення та аналізу інформації з різних джерел;
- ЗК11 Здатність приймати обґрунтовані рішення.

Спеціальні компетентності:

СК2 Здатність до виявлення статистичних закономірностей недетермінованих явищ, застосування методів обчислювального інтелекту, зокрема статистичної, нейромережевої та нечіткої обробки даних, методів машинного навчання та генетичного програмування тощо;

СК7 Здатність застосовувати теоретичні та практичні основи методології та технології моделювання для дослідження характеристик і поведінки складних об'єктів і систем, проводити обчислювальні експерименти з обробкою й аналізом результатів;

СК11 Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач.

3. Обсяг курсу

Вид заняття	150 годин / 5 кредитів ECTS		
	форма навчання		
	денна	заочна	дистанційна
лекції	28		
практичні заняття	14		
лабораторні роботи	14		
самостійна робота	60		
консультації	4		
підсумковий контроль	30		

4. Технічне обладнання і програмне забезпечення

Технічне обладнання: комп'ютер/ноутбук.

Програмне забезпечення:

- python 3.8 (<https://www.python.org/downloads/>);
- Jupyter Notebooks (<https://jupyter.readthedocs.io/en/latest/install.html>).

5. Політики курсу

Студенти мають дотримуватись кодексу академічної доброчесності:
https://www.krok.edu.ua/download/nakazi/2018-10-18_kodeks-akademichnoi-dobrochesnosti.pdf

6. Схема курсу

Тиждень / кіль-ть годин	Тема	Форма діяльності	Завдання / кіль-ть годин	Термін виконання
Тиждень #1 / 8 годин	Масиви та матриці. Створення та дослідження масивів.	лекція	Тестування / 2	1 тиждень
	Використання бібліотеки NumPy для обробки великих багатовимірних (числових) масивів та матриць	практичне заняття	Вирішення практичних завдань / 2	
	Математичні функції бібліотеки NumPy для роботи з масивами	самостійна робота	Відповіді на контрольні запитання / 4	

Тиждень #2 / 8 годин	Індексція 2D, 3D-масивів, стек та відбір масивів (Selection, Slicing)	лекція	Тестування / 2	1 тиждень
	Операції векторизації та "Broadcasting" з використанням NumPy	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Стандартизація масивів	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #3 / 8 годин	Структури даних: Series і DataFrame. Підготовка та очищення даних.	лекція	Тестування / 2	1 тиждень
	Обробка даних за допомогою бібліотеки Pandas	практичне заняття	Вирішення практичних завдань / 2	
	Фільтрація та сортування в Pandas	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #4 / 8 годин	Описова статистика категоріальних та кількісних даних	лекція	Тестування / 2	1 тиждень
	Описова статистика в Pandas	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Робота з відсутніми даними	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #5 / 8 годин	Візуалізація даних: основні графіки, діаграми розсіювання, Boxplot та інші	лекція	Тестування / 2	1 тиждень
	Використання бібліотек Matplotlib & Seaborn для візуалізації даних	практичне заняття	Вирішення практичних завдань / 2	
	Множинна вісь (Multiple axis)	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #6 / 8 годин	Основи одновимірної статистики. Точкові оцінки	лекція	Тестування / 2	1 тиждень
	Основи статистики з Numpy: Середнє, Дисперсія, Стандартне відхилення, Коваріація, Кореляція, Стандартна помилка (SE)	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Властивості точкових оцінок	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #7 / 8 годин	Основні розподіли: Нормальний розподіл, Розподіл Chi-Square, F-розподіл Фішера, t-розподіл Стюдента	лекція	Тестування / 2	1 тиждень
	Побудова графіків функцій: щільність ймовірності (pdf), функція розподілу ймовірностей (cdf), Survival function.	практичне заняття	Вирішення практичних завдань / 2	
	Дослідження зв'язку між основними розподілами	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #8 / 8 годин	Перевірка гіпотез статистичних гіпотез	лекція	Тестування / 2	1 тиждень
	Одновибірковий t-критерій (One sample t-test)	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Розрахунок p-value	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #9 / 8 годин	Тестування парних асоціацій. Кореляція Пірсона: перевірка зв'язку між двома кількісними змінними	лекція	Тестування / 2	1 тиждень
	Двовимірний (Student) t-test: порівняння двох середніх величин	практичне заняття	Вирішення практичних завдань / 2	
	Welch's t-тест	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #10 / 8 годин	Дисперсійний аналіз (ANOVA)	лекція	Тестування / 2	1 тиждень
	Критерій хі-квадрат для перевірки гіпотези незалежності частот в крос-	лабораторна робота	Звіт з виконання лабораторної роботи / 2	

	таблиці			
	Непараметричний тест парних асоціацій (кореляція Спірмена, тест Wilcoxon, Mann–Whitney U test)	самостійна робота	Відповіді на контрольні запитання / 4	
Тиждень #11 / 9 годин	Лінійна регресія: проста та множинна	лекція	Тестування / 2	1 тиждень
	Лінійні моделі з використанням бібліотеки Statsmodels	практичне заняття	Вирішення практичних завдань / 2	
	Корекція Бонферроні для множинних порівнянь	самостійна робота	Відповіді на контрольні запитання / 5	
Тиждень #12 / 9 годин	Багатовимірна статистика. Матриці коваріації, кореляції, часткової кореляції, точності	лекція	Тестування / 2	1 тиждень
	Багатовимірний нормальний розподіл	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Евклідова норма і норма Махаланобіса	самостійна робота	Відповіді на контрольні запитання / 5	
Тиждень #13 / 9 годин	Динамічний ряд (Time series). Стаціонарність. Передискретизація, згладжування, ковзне середнє, перша різниця.	лекція	Тестування / 2	1 тиждень
	Аналіз часових рядів отриманих на Google Trends	практичне заняття	Вирішення практичних завдань / 2	
	Декомпозиція часового ряду на тренд, сезонність та залишки (residuals)	самостійна робота	Відповіді на контрольні запитання / 5	
Тиждень #14 / 9 годин	Автокореляція. Функція автокореляції (ACF). Функція часткової автокореляції (PACF)	лекція	Тестування / 2	1 тиждень
	Прогнозування часових рядів за допомогою Python з використанням моделей ARMA	лабораторна робота	Звіт з виконання лабораторної роботи / 2	
	Модель ARIMA	самостійна робота	Відповіді на контрольні запитання / 5	

7. Система оцінювання та вимоги

Загальна система оцінювання курсу	Оцінювання проводиться за 100-бальною накопичувальною шкалою. Бали накопичуються за наступним співвідношенням: • тренувальні тести – 10% семестрової оцінки; • лабораторні роботи – 30% семестрової оцінки; • практичні роботи – 30% семестрової оцінки; • іспит / індивідуальний проект – 30% семестрової оцінки.
Практичні заняття / лабораторні роботи	Здача вирішених практичних завдань і звітів з лабораторних робіт повинна здійснюватися відповідно до встановлених дедлайнів. Виконані завдання завантажуються студентом у відповідні модулі на платформі Moodle.
Умови допуску до підсумкового контролю	Виконання всіх лабораторних робіт є обов'язковою умовою для допуску до складання іспиту

8. Рекомендована література до курсу

Базовий підручник

Edouard Duchesnay, Tommy Lofstedt, Feki Younes. Statistics and Machine Learning in Python. Engineering school. France. 2020. <https://duchesnay.github.io/pystatsml/>

Допоміжна література

1. Шумейко А. А. Интеллектуальный анализ данных (Введение в Data Mining) : учеб. пособ. / А. А. Шумейко, С. Л. Сотник. – Днепропетровск: Белая Е. А., 2012. – 212 с.
<http://pzs.dstu.dp.ua/Data/dm.pdf>
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning.
<https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6062a083acbf82c7195b27d/1617076404560/ISLR%2BSeventh%2BPrinting.pdf>

Онлайн-ресурси

1. Understanding and Visualizing Data with Python
<https://www.coursera.org/learn/understanding-visualization-data>
2. Inferential Statistical Analysis with Python <https://www.coursera.org/learn/inferential-statistical-analysis-python>
3. Fitting Statistical Models to Data with Python <https://www.coursera.org/learn/fitting-statistical-models-data-python>

До уваги студентів: усі навчально-методичні матеріали (презентації лекцій, тренувальні тести, тренувальні практичні завдання, завдання тощо) подані в Moodle за посиланням <https://dist1.krok.edu.ua/course/view.php?id=1273> .