



Кам'янець-Подільський національний університет
імені Івана Огієнка
Фізико-математичний факультет
Кафедра комп'ютерних наук

СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Інженерія даних з Python

1. Загальна інформація про курс. Викладачі

Іванюк Віталій Анатолійович, доктор технічних наук, доцент, завідувач кафедри комп'ютерних наук
E-mail: wivanyuk@kpmu.edu.ua
Профайл: <https://cs.kpmu.edu.ua/2019/11/04/ivaniuk-vitalij-anatolijovych/>
Консультації: щочетверга – 16.00-18.00 (очно та онлайн)

Пилипюк Тетяна Михайлівна, доцент, доцент кафедри комп'ютерних наук
E-mail: pylypyuk.tetiana@kpmu.edu.ua
Профайл: <https://cs.kpmu.edu.ua/2019/10/30/pylypiuk-tetiana-mykhajlivna/#more-649>

Мова викладання: українська

2. Анотація до курсу

Курс присвячений фундаментальним процесам інженерії даних, які є необхідною передумовою для будь-якого успішного проєкту в галузі машинного навчання чи аналізу даних. Студенти отримають системні знання та практичні навички з побудови відтворюваних пайплайнів підготовки даних мовою Python.

Основна увага приділяється програмній взаємодії з джерелами даних (зокрема, через SQL), методам очищення "сирих" даних та їх глибокої трансформації за допомогою бібліотеки Pandas. Ключовим елементом курсу є вивчення інженерії ознак (Feature Engineering) — процесу вибору, створення та перетворення предикторів, що є критичним для максимізації якості аналітичних моделей.

3. Мета і завдання курсу

Метою курсу є формування у студентів системних знань та практичних навичок щодо процесів інженерії даних (Data Engineering) та інженерії ознак (Feature Engineering). Курс зосереджений на використанні мови Python та її ключових бібліотек (зокрема, Pandas) для забезпечення повного циклу підготовки даних до подальшого аналізу та машинного навчання.

Завдання курсу:

- Опанувати інструментарій Python (зокрема, бібліотеку Pandas) для ефективної маніпуляції, агрегації та трансформації структурованих даних.
- Сформувані навички програмної взаємодії з реляційними базами даних (БД) та застосування SQL для вибірки, фільтрації та інтеграції даних.
- Навчити методам очищення "сирих" даних (data cleaning), виявлення аномалій, обробки пропущених значень та приведення даних до єдиного формату.
- Надати ґрунтовні знання з інженерії ознак (Feature Engineering): методів створення, вибору та перетворення предикторів, які підвищують якість моделей машинного навчання.
- Розвинути вміння проєктувати та реалізовувати відтворювані пайплайни підготовки даних (data pipelines) для вирішення практичних аналітичних задач.

4. Результати навчання

Після успішного завершення курсу студент зможе:

1. Програмно взаємодіяти з реляційними базами даних, використовуючи SQL та Python для вибірки, фільтрації та агрегації даних, необхідних для аналізу.
2. Застосовувати ключові методи бібліотеки Pandas для ефективної маніпуляції (трансформації, об'єднання, агрегації) та очищення "сирих" наборів даних (обробка пропусків, аномалій, дублікатів).

3. Володіти основними техніками інженерії ознак (Feature Engineering), свідомо створювати (генерувати) та перетворювати (кодувати) предиктори для підвищення якості прогнозних моделей.
4. Проєктувати та реалізовувати відтворювані пайплайни підготовки даних (data pipelines), що перетворюють "сирі" дані на готовий до аналізу, очищений набір ознак.

Писати відтворювані R-скрипти, впевнено використовуючи базовий синтаксис, функції та ключові структури даних

Імпортувати, очищувати та маніпулювати "сирими" даними з різноманітних джерел, готуючи їх для аналізу.

Інтегрувати R з реляційними базами даних: програмно підключатися до БД та виконувати SQL-запити для отримання, фільтрації та зміни даних.

"Розповідати історію" за допомогою даних, створюючи інформативні та готові до публікації візуалізації для ефективної комунікації результатів аналізу.

Розробляти повний аналітичний пайплайн: від підключення до джерела даних (файлу чи БД) до фінального звіту зі скриптами та візуалізаціями, які дають чіткі відповіді на поставлені запитання.

5. Формат курсу

Очний курс з елементами дистанційного навчання в системі Moodle.

6. Обсяг і ознаки курсу

Найменування показників	Характеристика навчальної дисципліни
Кількість кредитів ЄКТС	3/4
Загальний обсяг годин	90/120
Кількість годин навчальних занять	30/40
Лекційні заняття	12
Практичні заняття	0
Семінарські заняття	0
Лабораторні заняття	18/28
Самостійна та індивідуальна робота	60/80
Форма підсумкового контролю	Залік

7. Пререквізити і кореквізити курсу

Перелік дисциплін, які бажано, але не обов'язково, вивчити раніше: Програмування, Об'єктно-орієнтоване програмування, Організація баз даних та знань, Теорія ймовірностей та математична статистика, Інтелектуальний аналіз даних, Основ програмування мовою Python.

8. Технічне й програмне забезпечення, обладнання

Обладнання: ПК (Windows/macOS/Linux) з доступом до Інтернету.

Основне ПЗ: Python 3.9+, Git.

Середовище: VS Code (або JupyterLab).

Бібліотеки (Інженерія даних): Pandas, SQLAlchemy (та драйвери БД, напр., psycopg2)

9. Політика курсу

Увесь навчальний контент розміщено в модульному середовищі навчання К-ПНУ імені Івана Огієнка – moodle. Підготовка та виконання завдань лабораторних робіт є обов'язковими для кожного студента.

Академічна доброчесність. Очікується, що роботи студентів будуть їх оригінальними дослідженнями чи міркуваннями. Відсутність посилань на використані джерела, фабрикування джерел списування, втручання в роботу інших студентів становлять приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в роботі студента є підставою для її незарахування викладачем, незалежно від масштабів плагіату чи обману.

Відвідання занять. Очікується, що всі студенти відвідають усі лекції та лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати всі заняття. Виконання усіх лабораторних робіт є обов'язковим. Студенти зобов'язані дотримуватися термінів виконання усіх видів робіт, передбачених курсом.

Форми поточного та підсумкового контролю. Поточний контроль реалізується на лабораторних заняттях. Підсумковий контроль не здійснюється, а виставляється у формі заліку на основі поточного контролю.

10. Програма курсу

Модуль 1. Основи Python для Data Science та Pandas

- **Тема 1. Вступ до Data Engineering.**
 - **Зміст:** Роль інженерії даних. Огляд стеку Python (Pandas, SQLAlchemy). Встановлення середовища (Anaconda, Jupyter).
- **Тема 2. Pandas як основний інструмент.**
 - **Зміст:** Ключові структури: Series та DataFrame. Читання та запис даних (CSV, Excel). Базові операції: індексація, вибірка (loc, iloc), фільтрація.

Модуль 2. Доступ до даних та SQL

- **Тема 3. SQL для аналітика.**
 - **Зміст:** Огляд SELECT, WHERE, GROUP BY, HAVING.
- **Тема 4. Об'єднання даних (JOINS).**
 - **Зміст:** INNER, LEFT, RIGHT, OUTER JOINS. Підзапити.
- **Тема 5. Інтеграція SQL та Pandas.**
 - **Зміст:** Використання SQLAlchemy та Pandas (read_sql) для програмного доступу до баз даних (напр., SQLite, PostgreSQL) з Python.

Модуль 3. Очищення та трансформація даних

- **Тема 6. Дослідницький аналіз даних (EDA) та "Очищення".**
 - **Зміст:** Методи describe(), info(). Виявлення та обробка пропущених значень (.fillna(), .dropna()). Робота з дублікатами.
- **Тема 7. Трансформація даних.**
 - **Зміст:** Зміна типів даних (.astype()). Робота з рядками (.str.*) та датами (datetime). Використання .apply() та .map().
- **Тема 8. Агрегація та зведені таблиці.**
 - **Зміст:** groupby() в Pandas. Створення зведених таблиць (pivot_table), об'єднання (merge, concat).

Модуль 4. Інженерія ознак (Feature Engineering)

- **Тема 9. Основи Feature Engineering.**
 - **Зміст:** Що таке ознака (feature)? Важливість FE.
- **Тема 10. Робота з категоріальними даними.**
 - **Зміст:** Кодування: Label Encoding, One-Hot Encoding.
- **Тема 11. Робота з числовими даними.**
 - **Зміст:** Масштабування (Normalization, Standardization). Дискретизація (Binning). Створення поліноміальних та інтерактивних ознак.
- **Тема 12. Фінальний проєкт (Data Pipeline).**
 - **Зміст:** Побудова повного, відтворюваного скрипту (пайплайну), який бере "сирі" дані, очищує їх, генерує нові ознаки та зберігає фінальний "чистий" набір даних.

11. Система оцінювання та вимоги

Курс складається з 1-го навчального (змістовного) модуля.

Поточний контроль полягає в перевірці теоретичних знань та практичних умінь і навичок під час лабораторних занять.

Результати перевірки самостійної роботи, в тому числі виконання домашніх завдань, студентів також входять до поточного контролю. Максимальний бал оцінки поточної успішності студентів на практичних заняттях приймається рівним 12.

Підсумковий контроль проводиться у формі заліку.

Рейтингова оцінка студентам виставляється відповідно до Положення про рейтингову систему оцінювання навчальних досягнень здобувачів вищої освіти Кам'янець-Подільського національного університету імені Івана Огієнка після проведення всіх навчальних занять та контрольних заходів з навчальної дисципліни.

Студенти, які мають академічну заборгованість за результатами підсумкового контролю, зобов'язані ліквідувати її в терміни, визначені графіком ліквідації академічної заборгованості.

12. Рекомендована література

1. Wes McKinney. *Python for Data Analysis* (3rd Edition). O'Reilly, 2022.
2. Alan Beaulieu. *Learning SQL* (3rd Edition). O'Reilly, 2020.
3. Alice Zheng & Amanda Casari. *Feature Engineering for Machine Learning*. O'Reilly, 2018.
4. Офіційна документація Pandas (pandas.pydata.org/docs/)