



МОВИ ПРОГРАМУВАННЯ R ТА PYTHON В СТАТИСТИЧНИХ ОБЧИСЛЕННЯХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	<i>Другий (магістерський)</i>
Галузь знань	<i>C - соціальні науки, журналістика, інформація та міжнародні відносини</i>
Спеціальність	<i>C5 Соціологія</i>
Освітня програма	<i>Аналітика соціальних даних</i>
Статус дисципліни (код)	<i>Нормативна, Цикл професійної підготовки</i>
Форма навчання	<i>очна(денна)</i>
Рік підготовки, семестр	<i>1 курс, весняний семестр</i>
Обсяг дисципліни	<i>4 кредити ЕКТС, 120 годин лекції 16 год., практичні заняття 30 год., самостійна робота 74 год.</i>
Семестровий контроль/ контрольні заходи	<i>Залік, МКР, РР</i>
Розклад занять	https://schedule.kpi.ua/ <i>1 год. лекційних занять та 2 год. комп'ютерних практикумів на тиждень</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	Лектор: к.ф.-м.н., доцент, Пишнограєв Іван Олександрович, pyshnograiev@wdc.org.ua Комп. практикуми: к.ф.-м.н., доцент, Пишнограєв Іван Олександрович
Розміщення курсу	Google classroom https://classroom.google.com/c/NzQ4Mzc0MTAwMTIx?cjc=jr4j6zc

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Дисципліна є нормативною в освітній програмі. Вивчення навчальної дисципліни націлено на формування, розвиток та закріплення у здобувачів таких загальних та фахових компетентностей:

ЗК 01 Здатність до абстрактного мислення, аналізу та синтезу,

ФК 13 Здатність використовувати статистичний пакет SPSS та мови програмування R та Python для аналізу соціальних даних.

Внаслідок вивчення курсу студент повинен бути здатний продемонструвати такий програмний результат навчання ОНП:

ПРН 04 Застосовувати наукові знання, соціологічні та статистичні методи, цифрові технології, спеціалізоване програмне забезпечення для розв'язування складних задач соціології та суміжних галузей знань,

ПРН 14 Застосовувати мови програмування R та Python для аналізу соціальних даних.

У кінці вивчення курсу студент повинен **знати**:

- особливості роботи з мовами програмування R та Python;
- методи обробки та аналізу даних;
- засоби обробки, зберігання та аналізу даних;

вміти:

- аналізувати дані за допомогою мов програмування R та Python;
- створювати скрипти та програми для аналізу соціальних даних.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Дисципліна базується на знаннях та навичках суміжних дисциплін, що вивчаються на попередньому семестрі та освітньому рівні. Дана дисципліна передує ОК ПО 05 «Технології OSINT в соціологічних дослідженнях» та може являтися однією з головних складових магістерської дисертації.

3. Зміст навчальної дисципліни:

Розділ 1. Використання мови R для розв'язання задач аналізу даних

Тема 1.1. Основи мови R.

1. Характеристики мови R та середа програмування;
2. Базові типи даних та дії над ними, написання елементарних скриптів;
3. Основні структури даних.
4. Основні алгоритмічні елементи в мові R.
5. Робота з джерелами даних.

Тема 1.2. Візуалізація даних за допомогою мови R.

1. Вступ до візуалізації даних, основні види графіків;
2. Просунуті інструменти візуалізації, побудова мап;
3. Створення дашбордів.

Тема 1.3. Аналіз даних в R.

1. Постановка задачі, основні етапи;
2. Способи обробки даних різних типів, їх форматування, нормалізація тощо;
3. Попередній аналіз даних, проведення кореляційно-регресійного аналізу;
4. Методи прогнозування, побудова регресійних моделей;
5. Методи оцінки моделей.

Тема 1.4. Приклади використання R для аналізу соціально-економічних даних.

Розділ 2. Використання мови Python для розв'язання задач аналізу даних

Тема 2.1. Основи мови Python.

1. Характеристики мови R та середа програмування;
2. Базові типи даних та дії над ними, написання елементарних скриптів;
3. Основні структури даних.
4. Основні алгоритмічні елементи в мові Python.
5. Класи та об'єкти (оглядово).

Тема 2.2. Основні бібліотеки мови Python для роботи з джерелами даних.

1. Використання Pandas для роботи з даними;
2. Використання NumPy для роботи зі структурами даних.

Тема 2.3. Розвідувальний аналіз даних.

1. Описова статистика;
2. Кореляційно-регресійний аналіз.

Тема 2.4. Побудова моделей з використанням Python.

1. Лінійна та багатофакторна лінійна регресія;
2. Поліноміальна регресія.
3. Оцінка якості моделей

Тема 2.5. Приклади використання Python для аналізу соціально-економічних даних.

Перелік комп'ютерних практикумів:

1. Створення скриптів на мові R для аналізу та візуалізації даних.
2. Розв'язання задач аналізу соціальних даних з використанням мови R.
3. Створення скриптів на мові Python для аналізу та візуалізації даних.
4. Розв'язання задач аналізу соціальних даних з використанням мови Python.

3. Навчальні матеріали та ресурси

Базова:

1. *R for Data Science* by Hadley Wickham, Garrett Grolemund [en]. URL: <https://r4ds.had.co.nz/>.
2. *Advanced R* by Hadley Wickham [en]. URL: <https://adv-r.hadley.nz/index.html>
3. Майборода Р.Є Комп'ютерна статистика. Професійний старт. Навчальний посібник. Київський університет», 2020. – 482 с. <http://probability.univ.kiev.ua/userfiles/mre/compsta1.pdf>
4. Методи і моделі інтелектуального аналізу даних. Практикум [Електронний ресурс] : навчальний посібник для студентів, які навчаються за спеціальністю 122 «Комп'ютерні науки», освітньої програми «Системи і методи штучного інтелекту» / Н. І. Недашківська ; КПІ ім. Ігоря Сікорського. – Київ : КПІ ім. Ігоря Сікорського, 2019. – 71 с. <https://ela.kpi.ua/handle/123456789/53764>
5. Цеслів, О. В. Програмування для аналітичних досліджень [Електронний ресурс] : навчальний посібник для здобувача ступеня бакалавра за освітньою програмою Економічна аналітика, спеціальності 051 Економіка. Електронне мережеве видання / О. В. Цеслів ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 1,55 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2024. – 238 с. <https://ela.kpi.ua/handle/123456789/66102>.

Допоміжна:

1. Аналіз даних. Лабораторний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 113 «Прикладна математика» / Н. М. Кукуль, А. Ю. Шелестов, С. А. Тарасенко, Г. О. Яйлимова ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 582.97 Кбайт). – Київ: КПІ ім. Ігоря Сікорського, 2022. – 28 с. <https://ela.kpi.ua/handle/123456789/50425>
2. Новотарський, М. А. Основи програмування алгоритмічною мовою Python [Електронний ресурс] : навч. посіб. для студ. освітньої програми «Комп'ютерні системи та мережі» спеціальності 123 «Комп'ютерна інженерія» / М. А. Новотарський ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 17.93 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 701 с. <https://ela.kpi.ua/handle/123456789/49913>
3. *Data Mining Tutorial*. <https://www.geeksforgeeks.org/data-mining/>
4. *Python Data Mining Quick Start Guide*, published by Packt. <https://github.com/PacktPublishing/Python-Data-Mining-Quick-Start-Guide>
5. *R програмування* // Електронний ресурс. Режим доступу: <https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovaniye/r-programmirovaniye>
6. *Virtualization Technology* // Електронний ресурс. Режим доступу: <https://www.sciencedirect.com/topics/computer-science/virtualization-technology>
7. *The R Manuals*. <https://cran.r-project.org>
8. *Practical Data Mining with Python*. <https://dzone.com/refcardz/data-mining-discovering-and>

Навчальний контент

4. Методика опанування навчальної дисципліни (освітнього компонента)

У гугл-класі будуть викладатись завдання з детальними інструкціями та необхідним матеріалом, які необхідно вчасно виконувати.

Розділ 1. Використання мови R для розв'язання задач аналізу даних

Лекція 1. Основи мови R: синтаксис та структури даних

Історія та філософія мови R. Огляд середовища розробки RStudio. Поняття пакетів (packages) та репозиторію CRAN. Базові типи даних: numeric, integer, character, logical. Векторна арифметика. Основні структури даних: вектори, матриці, списки (lists) та датафрейми (data frames). Керуючі конструкції (if-else, цикли) та написання власних функцій. Імпорт та експорт даних (CSV, Excel).

Завдання на СРС:

1. Встановити R та RStudio, налаштувати робочу директорію.

2. Створити скрипт, що завантажує зовнішній файл даних, перевіряє типи змінних та виводить базову статистику (summary).

Лекція 2. Візуалізація даних в екосистемі R

Графічна система Base R vs. ggplot2. Концепція «Grammar of Graphics». Побудова основних типів графіків: гістограми, діаграми розсіювання, boxplots. Налаштування естетики (aesthetics), геометрій (geoms) та фасетування. Робота з геопросторовими даними: пакети sf та leaflet для побудови інтерактивних мап. Вступ до створення веб-додатків та дашбордів за допомогою R Shiny.

Завдання на CPC:

1. Ознайомитися з галереєю графіків R Graph Gallery.
2. Побудувати складний графік у ggplot2, використавши мінімум три шари (дані, геометрія, тема) та групування за кольором.

Лекція 3. Методологія аналізу даних в R

Концепція «Tidy Data» та використання пакетів dplyr і tidyr для маніпуляцій з даними (фільтрація, агрегація, з'єднання таблиць). Попередній аналіз (EDA): виявлення викидів, перевірка розподілу. Кореляційний аналіз: розрахунок коефіцієнтів Пірсона/Спірмена, побудова корелограм. Побудова лінійної регресії (функція lm). Діагностика моделі та інтерпретація результатів. Метрики якості: R-squared, RMSE.

Завдання на CPC:

1. Вивчити «чім-лист» по роботі з dplyr.
2. Провести очищення "брудного" датасету: перейменувати колонки, замінити пропущені значення середнім або медіаною.

Лекція 4. R в соціально-економічних дослідженнях

Специфіка роботи з даними опитувань (зважування вибірки, пакет survey). Аналіз категоріальних даних та таблиць спряженості. Робота зі шкалами Лайкерта. Використання API для отримання економічних даних (наприклад, пакети WDI для даних Світового банку або eurostat). Приклади візуалізації демографічних змін та економічних трендів.

Завдання на CPC:

1. Знайти пакет R, що надає доступ до відкритих державних даних (Open Data) або даних міжнародних організацій.
2. Завантажити набір соціально-економічних показників для однієї країни за останні 10 років.

Розділ 2. Використання мови Python для розв'язання задач аналізу даних

Лекція 5. Вступ до Python для аналізу даних

Python як універсальна мова програмування. Середовища Jupyter Notebook та Google Colab. Базові типи даних: int, float, str, bool. Структури даних: списки (lists), кортежі (tuples), словники (dictionaries), множини (sets). Основи алгоритмізації: цикли (for, while), умовні оператори, генератори списків (list comprehensions). Основи об'єктно-орієнтованого програмування: поняття класу, об'єкта, методу та атрибуту. Завдання на CPC:

1. Написати функцію на Python, яка обробляє список чисел (наприклад, фільтрує парні числа та підносить їх до квадрату).

2. Ознайомитися з відмінностями в індексації масивів у R (з 1) та Python (з 0).

Лекція 6. Бібліотеки NumPy та Pandas: фундамент Data Science

Бібліотека NumPy: поняття ndarray, векторизація обчислень, broadcasting. Бібліотека Pandas: об'єкти Series та DataFrame. Завантаження даних з різних джерел. Індексція, зрізи (slicing) та фільтрація даних (loc, iloc). Групування та агрегація даних (groupby, pivot_table). Обробка пропущених значень та дублікатів.

Завдання на СРС:

1. Виконати операцію merge (об'єднання) двох датафреймів за спільним ключем.
2. Порівняти швидкість виконання математичних операцій зі стандартними списками Python та масивами NumPy.

Лекція 7. Розвідувальний аналіз даних (EDA) засобами Python

Описова статистика за допомогою Pandas (describe, value_counts). Візуалізація розподілів та залежностей: бібліотеки Matplotlib та Seaborn. Побудова гістограм, boxplot, heatmap для кореляційної матриці. Статистичні гіпотези: бібліотека scipy.stats (t-тест, перевірка нормальності розподілу).

Завдання на СРС:

1. Побудувати матрицю кореляції для багатовимірного датасету та візуалізувати її за допомогою теплової карти (heatmap).
2. Виявити та візуалізувати викиди в даних за допомогою діаграми розмаху ("ящик з вусами").

Лекція 8. Моделювання та машинне навчання в Python. Python в прикладних соціальних дослідженнях

Вступ до бібліотеки scikit-learn та statsmodels. Лінійна регресія: проста та множинна. Поліноміальна регресія як спосіб роботи з нелінійними зв'язками. Розбиття даних на навчальну та тестову вибірки (train_test_split). Проблема перенавчання (overfitting). Метрики оцінки якості регресійних моделей: MAE, MSE, R2.

Специфіка бібліотек для аналізу текстів (NLTK, spaCy) в соціології. Робота з часовими рядами в економіці (бібліотека statsmodels, декомпозиція трендів). Отримання даних із соціальних мереж та веб-скрейпінг (BeautifulSoup, Requests). Автоматизація збору статистики.

Завдання на СРС:

1. Побудувати модель лінійної регресії для прогнозування ціни на нерухомість (або іншого показника) на тестовому наборі даних.
2. Порівняти метрики якості для лінійної та поліноміальної моделей.
3. Написати простий скрипт для отримання даних з відкритого API (наприклад, погода, курси валют або новини).
4. Інтерпретувати результати регресійної моделі в контексті соціального явища (наприклад, вплив освіти на рівень доходу).

Практичні заняття (комп'ютерні практикуми)

Комп'ютерний практикум 1. Створення скриптів на мові R для аналізу та візуалізації даних (6 год.)

Мета: опанувати базовий синтаксис R, навчитися імпортувати «сирі» дані, приводити їх до єдиного вигляду та створювати інформативні візуалізації.

1. Імпорт CSV-файлу. Використання пакету `dplyr` для фільтрації записів, вибору колонок та створення нових змінних.

Групування даних за категоріальною ознакою та розрахунок зведених статистик (середнє, медіана, стандартне відхилення).

2. Використання `ggplot2`. Побудова гістограми розподілу та діаграми розсіювання з лінією тренду.
3. Збереження отриманих результатів у новий файл та графіків у формат PNG/PDF.

Комп'ютерний практикум 2. Розв'язання задач аналізу соціальних даних з використанням мови R (8 год.)

Мета: застосувати статистичний апарат R для перевірки гіпотез та пошуку взаємозв'язків у реальних соціальних даних.

План

1. Побудова кореляційної матриці для набору соціальних індикаторів (наприклад, рівень безробіття та рівень злочинності). Візуалізація корелограми.
2. Побудова лінійної моделі. Аналіз коефіцієнтів, *p-value* та *R-squared*. Перевірка залишків моделі на нормальність.
3. Відображення показників на карті регіонів України або світу (хлороплет-карта) за допомогою бібліотек `sf` або `tmap`.
4. Формування короткого звіту з висновками в R Markdown.

Комп'ютерний практикум 3. Створення скриптів на мові Python для аналізу та візуалізації даних (8 год.)

Мета: навчитися використовувати бібліотеки `Pandas` та `Matplotlib/Seaborn` для маніпуляцій з табличними даними та їх графічного представлення.

План

1. Створення `DataFrame`, завантаження даних. Використання методів `.info()`, `.describe()`, `.value_counts()`. Робота з індексами та зрізами.
2. Знаходження та заміна пропущених значень. Кодування категоріальних змінних (`get_dummies`).
3. Побудова `subplot` (сітки графіків) за допомогою `Matplotlib`. Створення `boxplot` у `Seaborn` для порівняння розподілів у різних групах.
4. Перетворення колонки дати у формат `datetime`, побудова графіка динаміки показника у часі.

Комп'ютерний практикум 4. Розв'язання задач аналізу соціальних даних з використанням мови Python (6 год.)

Мета: реалізувати повний цикл побудови прогностної моделі на основі соціально-економічних даних.

План

5. Вибір цільової змінної та предикторів. Розуміння на `X_train`, `X_test`, `y_train`, `y_test`.
6. Навчання моделі лінійної регресії (`LinearRegression` з `scikit-learn`). Навчання моделі поліноміальної регресії для порівняння.
7. Розрахунок помилок прогнозування (MSE, MAE) на тестовій вибірці. Порівняння ефективності моделей.
8. Аналіз ваг (коефіцієнтів) моделі. Які соціальні фактори найбільше впливають на цільовий показник. Візуалізація лінії регресії на фоні реальних даних.

Модульна контрольна робота (2 год.)

9. Самостійна робота студента

Індивідуальні завдання складаються з підготовки до комп'ютерних практикумів, опрацюванні лекційного матеріалу та виконання розрахункової роботи.

Розрахункова робота полягає у створенні особистого міні-проєкту по аналізу соціальних даних з використанням будь-якої мови програмування. Тему студент обирає самостійно та затверджує з викладачем. За необхідності викладач може запропонувати тему на вибір.

Самостійна робота здобувача включає:

- підготовка до аудиторних занять – 54 год;
- розрахункова робота – 10 год;
- підготовка до МКР – 4 год;
- підготовка до заліку – 6 год.

Загалом – 74 год.

Політика та контроль

10. Політика навчальної дисципліни (освітнього компонента)

Усі роботи студенти мають прикріплювати в особистому кабінеті гугл-класу. Дедлайни кожного завдання позначені в завданнях у гугл-класі. Роботи мають бути виконані з дотриманням академічної доброчесності. Політика та принципи академічної доброчесності, етична поведінка студентів визначені у Кодексі честі <https://kpi.ua/code>. Лектор може запропонувати студентам пройти онлайн-курси на платформі Coursera. Також сертифікати цих курсів можуть бути частково зараховані згідно до Положення.

Тематика робіт спрямована на поглиблення засвоєного матеріалу лекцій. На заняттях комп'ютерного практикуму розв'язуються задачі та вправи по темах лекції.

11. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

Семестровий контроль: **залік**.

Семестровий рейтинг студента з дисципліни виставляється лектором та складається з балів, що він отримує за:

- ~ виконання модульної контрольної;
- ~ виконання 4 комп'ютерних практикумів;
- ~ виконання розрахункової роботи.

Критерії нарахування балів за семестр:

- 1) Модульна контрольна робота оцінюється у 20 балів.
- 2) Кожен з практикумів оцінюється в 14 балів - $14 \times 4 = 56$ балів.
- 3) Розрахункова робота оцінюється в 24 балів.

Критерії нарахування балів за контрольні заходи:

- "відмінно": 95 -100% - здобувач виявив всебічні, систематичні та глибокі знання навчального матеріалу з дисципліни; продемонстрував уміння вільно виконувати всі завдання, передбачені програмою; засвоїв основну та додаткову літературу; проявив творчі здібності в розумінні, в логічному, чіткому, стислому та ясному трактуванні навчального матеріалу; засвоїв взаємозв'язок основних понять дисципліни, їх значення для подальшої професійної діяльності
- "дуже добре": 85-94% - здобувач виявив систематичні знання навчального матеріалу з дисципліни вище середнього рівня; продемонстрував уміння добре виконувати всі завдання, передбачені програмою, допустивши незначні помилки; засвоїв основну та додаткову літературу; засвоїв взаємозв'язок основних понять дисципліни, їх значення для подальшої професійної діяльності
- "добре": 75-84% - здобувач виявив загалом добрі знання навчального матеріалу при виконанні передбачених програмою завдань, але припустив ряд помітних помилок; засвоїв основну літературу; показав систематичний характер знань з дисципліни; здатний до їх самостійного використання та поповнення в процесі подальшої навчальної роботи і професійної діяльності
- "задовільно": 65-74% - здобувач виявив знання основного навчального матеріалу з дисципліни в обсязі, необхідному для подальшого навчання та майбутньої професійної діяльності; ознайомився з основною літературою; впорався з виконанням завдань, передбачених

програмою, але припустив значну кількість помилок або недоліків на запитання при співбесіді, тестуванні та при виконанні завдань тощо, принципів з яких може усунути самостійно

- “достатньо”: 60-64% - здобувач виявив знання основного навчального матеріалу з дисципліни в мінімальному обсязі, необхідному для подальшого навчання та майбутньої професійної діяльності; ; ознайомився з основною літературою; в основному виконав завдання, передбачені програмою, але припустив помилки у відповіді на запитання при співбесідах, тестуванні та при виконанні завдань тощо, які він може усунути лише під керівництвом та за допомогою викладача
- “незадовільно”: 30-54% - здобувач мав значні прогалини в знаннях основного навчального матеріалу; допускав принципові помилки при виконанні передбачених програмою завдань, але спроможний самостійно доопрацювати програмний матеріал і підготуватися для перездачі дисципліни
- “незадовільно”: 0-29% - здобувач не мав знань зі значної частини навчального матеріалу з дисципліни; допускав принципові помилки при виконанні більшості передбачених програмою завдань або не виконував ці завдання

Умовою першої атестації є поточний рейтинг не менше 30% запланованих балів за семестр. Умова другої атестації ~ поточний рейтинг не менше 60% запланованих балів.

Необхідною умовою допуску до заліку є зарахування розрахункової роботи та 36 балів семестрового рейтингу.

На залікову роботу виноситься 100 балів. Залік проводиться у вигляді письмової роботи, в якій три теоретичних питання та два практичних. Кожне завдання оцінюється в 20 балів за такими критеріями: «відмінно», повна відповідь, не менше 90% потрібної інформації, що виконана згідно з вимогами до рівня «умінь», (повне, безпомилкове розв’язування завдання) – 18-20 балів;

«добре», достатньо повна відповідь, не менше 75% потрібної інформації, що виконана згідно з вимогами до рівня «умінь» або є незначні неточності (повне розв’язування завдання з незначними неточностями) – 14-16 балів;

«задовільно», неповна відповідь, не менше 60% потрібної інформації, що виконана згідно з вимогами до «стереотипного» рівня та деякі помилки (завдання виконане з певними недоліками) – 12 балів;

«незадовільно», відповідь не відповідає умовам до «задовільно» – 0 балів.

Сума рейтингових балів, отриманих студентом протягом семестру, переводиться до підсумкової оцінки згідно з таблицею.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Бали:	Оцінка
100...95	Відмінно
94...85	Дуже добре
84...75	Добре
74...65	Задовільно
64...60	Достатньо
Менше 60	Незадовільно
не зараховано розрахункову роботу або менше 36	Не допущено

Робочу програму навчальної дисципліни (силабус):

Складено доцент, к.ф.-м.н., доцент Пишнограєв Іван Олександрович



Ухвалено кафедрою ШІ (протокол № 14 від 24.06.2025)

Погоджено Методичною комісією НН ІПСА (протокол № 7 від 25.06.2025)