Форма № Н-3.04

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КРЕМЕНЧУЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ МИХАЙЛА ОСТРОГРАДСЬКОГО
Навчально-науковий інститут електричної інженерії
та інформаційних технологій
Кафедра комп'ютерної інженерії та електроніки

«ЗАТВЕРДЖУЮ»
Проректор з науково-педагогічної та
методичної роботи
_____ Віктор КОСТІН
_____ 2024 року

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

«Інтелектуальний аналіз даних»

другого (магістерського) освітнього рівня
спеціальності 123 «Комп'ютерна інженерія»
освітньо-професійної програми «Комп'ютерна інженерія»

КРЕМЕНЧУК 2024

2

Робоча програма навчальної дисципліни «Інтелектуальний аналіз даних» розроблена на основі освітньо-професійної програми «Комп'ютерна інженерія» підготовки здобувачів вищої освіти освітнього ступеня «Магістр» за спеціальністю 123 «Комп'ютерна інженерія» та відповідних нормативних документів

Робочу програму розробив:
доцент кафедри КІЕ, к. т. н. Валерій СИДОРЕНКО
(Власне ім'я ПРІЗВИЩЕ)

Робочу програму обговорено та схвалено на засіданні випускової кафедри освітньо-професійної програми «Комп'ютерна інженерія» спеціальності 123 «Комп'ютерна інженерія»,

протокол № 6 від 23.02.2024

Гарант освітньої програми _____ Микола ГУЧЕНКО
(підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Завідувач кафедри _____ Андрій ПЕРЕКРЕСТ
(підпис) *(Власне ім'я ПРІЗВИЩЕ)*

Робочу програму обговорено та схвалено на засіданні науково-методичної ради інституту електричної інженерії та інформаційних технологій,

протокол № 5 від 23.02.24

Голова науково-методичної ради _____ Юрій ЗАЧЕПА
(підпис) *(Власне ім'я ПРІЗВИЩЕ)*

1. Опис навчальної дисципліни

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, спеціальність, освітньо-професійна програма, освітній ступінь</td><td>Характеристика навчальної дисципліни</td></tr>
<tr><td>денна форма</td></tr>
<tr><td>Кількість кредитів – 5</td><td>Галузь знань
12 «Інформаційні технології»</td><td>Вибіркова</td></tr>
<tr><td>Модулів – 1</td><td rowspan="2">Спеціальність
123 «Комп’ютерна інженерія»</td><td>Рік підготовки</td></tr>
<tr><td>Змістових модулів – 3</td><td>1-й</td></tr>
<tr><td>Індивідуальне науково-дослідне завдання РР
(КР, КП, РР, РГ, к/р)</td><td rowspan="3">Освітньо-професійна програма
«Комп’ютерна інженерія»</td><td>Семестр</td></tr>
<tr><td rowspan="2">Загальна кількість годин – 150</td><td>1-й</td></tr>
<tr><td>Лекції</td></tr>
<tr><td rowspan="8">Тижневих годин для денної форми навчання:
аудиторних – 3
самостійної роботи студента – 5</td><td rowspan="8">Освітній ступінь
«Магістр»</td><td>30 год.</td></tr>
<tr><td>Лабораторні</td></tr>
<tr><td>20 год.</td></tr>
<tr><td>Самостійна робота</td></tr>
<tr><td>100 год.</td></tr>
<tr><td>Вид контролю</td></tr>
<tr><td>диф. залік</td></tr>
</table>

Співвідношення кількості годин аудиторних занять до самостійної і індивідуальної роботи становить:

для денної форми навчання – 50/100=0,5.

2. Мета та завдання навчальної дисципліни

Мета: надати студентам знання і привити практичні навички з основ сучасних технологій інтелектуального аналізу даних.

Завдання: забезпечити системний підхід до розробки закінчених аналітичних рішень із застосуванням технології Data Mining на базі спеціалізованої мови програмування R та її фреймворків на рівні вимог, що висуваються на ринку праці до аплікантів на наступні юніорські позиції: аналітик даних (Data Analyst) та вчений по даним (Data Scientist).

У результаті вивчення навчальної дисципліни студент повинен

отримати досвід з компетентностей:

ЗК 2. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК 3. Здатність проводити дослідження на відповідному рівні.

ЗК 4. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

СК 2. Здатність представляти результати власних досліджень та/або розробок у вигляді презентацій, науковотехнічних звітів, статей і доповідей на науково-технічних конференціях..

набути навички та уміння:

ПРН 2. Знаходити необхідні дані, аналізувати та оцінювати їх.

ПРН 10. Здійснювати пошук інформації в різних джерелах для розв’язання задач комп’ютерної інженерії, аналізувати та оцінювати цю інформацію.

Згідно з вимогами освітньо-професійної програми студент повинен

знати:

– теоретичні основи технології Data Mining та засоби реалізації Data Science-проектів;

– базові структури даних та спосіб реалізації їх на мові Python;

– знати основні алгоритми роботи з даними: на рядках, на графах, оптимального кодування без втрат;

уміти:

– розробляти закінчені програмно-аналітичні рішення (так званий Data Science-проект) згідно зі стандартом CRISP DM на базі мови програмування R у середовищі IDE RStudio із застосуванням фреймворків RMarkdown, Quarto, Shiny та ін.

у рамках Data Science-проекту:

– виконувати імпорт-експорт даних з різних джерел, включаючи реляційні БД;

– маніпулювати даними на рівні інтерфейсу командного рядка на етапах розвідувального аналізу даних та підготовки даних до аналізу;

– виконувати візуалізацію даних і результатів аналізу із застосуванням табличних та графічних візуалізаторів;

– виконувати побудову моделей інтелектуального аналізу: кластеризації, класифікації, регресії, асоціативних правил на основі статистичного та машинного навчання;

– виконувати розгортання розробленого рішення у вигляді веб-застосунку із використанням клієнт-серверних технологій.

3. Програма навчальної дисципліни

Змістовий модуль 1 Базовий. Основи Data Science-проекту. Розвідувальний аналіз даних.

Тема 1. Введення у дисципліну

Що таке інтелектуальний аналіз даних (ІАД, Data Mining)? Задачі, які дозволяє вирішити ІАД в конкретних галузях. Приклади успішних проектів із застосуванням технології ІАД.

Що таке Data Science? Що таке Data Science-проект? Структура типового Data Science-проекту. Концепція грамотного програмування. Markdown і RMarkdown. Інсталяція R. Інсталяція RStudio. Створення RMarkdown-документу. Генерація електронного документу. Видавнича система Quarto

Тема 2. Маніпулювання даними

Мета і задачі маніпулювання даними. Структуровані, слабоструктуровані і неструктуровані дані. Імпорт даних. Імпорт даних з джерел різного формату. Імпорт даних з реляційних баз даних. Засоби маніпулювання даними.

Приведення даних до охайного вигляду. Що таке «охайні дані» (tidy data)? Існуючі засоби приведення даних до охайного вигляду. Обробка пропущених значень.

Трансформація даних. Засоби трансформації даних.

Тема 3. Розвідувальний аналіз даних. Візуалізація

Мета і задачі розвідувального аналізу даних. Питання (гіпотеза) як центральне поняття розвідувального аналізу даних. Варіація і коваріація – два загальні аспекти розвідувального аналізу даних в контексті аналізу числових і нечислових даних. Засоби розвідувального аналізу даних.

Зниження розмірності, мета і задачі, проектування і обирання ознак. Методи зниження розмірності: аналіз головних компонент (PCA), факторний

аналіз (FA), багатовимірне шкалювання (MS). Сегментація (кластеризація даних). Нелінійні методи зниження розмірності.

Візуалізація, як основний інструмент подання поточних і вихідних результатів аналізу даних. Засоби візуалізації.

Змістовий модуль 2 Просунутий рівень. Моделювання. Основні задачі, алгоритми та засоби

Тема 4. Моделювання. Класифікація

П’ять основних класів задач Data Mining.

Постановка задачі класифікації. Зв’язок задач класифікації та регресії. Статистичні моделі і моделі на основі машинного навчання. Логістична регресія. Лінійний дискримінантний аналіз. Байєсівській класифікатор. Алгоритми CART та C 4.5.

Засоби побудови моделей класифікації.

Тема 5. Моделювання. Регресія

Постановка задачі регресії. Статистичні підходи: лінійна і нелінійна регресія. Метод найменших квадратів (МНК). Пуассонівська регресія. Підходи на основі машинного навчання: нейронні мережі, градієнтний бустінг, дерева регресії та випадкові ліси (random forest).

Технологія побудови моделі регресії. Відбір і трансформація регресорів, поняття навчальної та тренувальної вибірок. Перевірка адекватності моделей, кросс-валідація.

Засоби збереження побудованих моделей для цілей практичного використання у майбутніх проектах. Серилізація.

Тема 6. Моделювання. Пошук асоціативних правил

У чому полягає задача пошуку асоціативних правил? Формальна постановка задачі. Транзакція. Метрики: підтримка, достовірність, ліфт. Засоби побудови, оцінки та візуалізації асоціативних моделей.

Тема 7. Моделювання. Аналіз часових рядів. ARIMA-моделі часового ряду. Прогнозування

Поняття часового ряду (ЧР). Стаціонарність і дифернціювання ЧР. Несезонні ARIMA-моделі: AR(1), AR(2), MA(1), MA(2), ARMA(1, 1). Сезонні моделі. Структурна і параметрична ідентифікація ARIMA-моделі. Перевірка адекватності. Метрики якості. Прогнозування на основі ARIMA-моделі.

Засоби побудови ARIMA-моделей.

Тема 8. Веб-скрепінг

Веб-скрепінг та його етапи: http-запит, отримання сторінки, парсинг сторінки. Пакети `RCurl`, `httr`, `rvest`. Основні методи пакета `rvest`. Отримання html-сторінки і знаходження потрібного елемента, XPathта CSS-селектори. Розбір елементів на складові частини.

Змістовий модуль 3. Засоби автоматизації розробки та розгортання Data Science-проекту

Тема 9. Створення дашбордів та розгортання Data Science-проекту

Створення дашбордів.

Існуючі засоби розгортання Data Science-проекту. Клієнт-серверні технології. Хмарні сервіси.

Фреймворки Shiny, RestRserve. Хмарний сервіс shinyapps.io.

4. Структура навчальної дисципліни

Назви змістових модулів і тем	Кількість годин				
	денна форма				
	усього	у тому числі			
		л.	п. р.	л. р.	с. р.
1	2	3	4	5	6
Модуль 1 (семестр 1)					
Змістовий модуль 1. Базовий. Основи Data Science-проекту. Розвідувальний аналіз даних					
Тема 1. Введення у дисципліну.	9	2	–	2	5
Тема 2. Маніпулювання даними.	9	2	–	2	5
Тема 3. Розвідувальний аналіз даних. Візуалізація. Зниження розмірності	16	4	–	2	10
Разом за змістовим модулем 1	34	8	–	6	20
Змістовий модуль 2. Просунутий рівень. Моделювання. Основні задачі та алгоритми					
Тема 4. Моделювання. Класифікація.	16	4	–	2	10
Тема 5. Моделювання. Регресія.	16	4	–	2	10
Тема 6. Моделювання. Пошук асоціативних правил	9	2	–	2	5
Тема 7. Моделювання. Аналіз часових рядів. ARIMA-моделі часового ряду. Прогнозування	23	6	–	2	15
Тема 8. Веб-скрепінг	9	2	–	2	5
Разом за змістовим модулем 2	73	18	–	10	45
Змістовий модуль 3. Засоби автоматизації розробки та розгортання Data Science-проекту					
Тема 9. Створення дашбордів та розгортання Data Science-проекту	11	4	–	2	5
Фінальний проєкт (РР)	22		–	2	20
Разом годин за змістовим модулем 3	33	4	–	4	25
Підсумковий контроль: диф. залік	10	–	–	–	10
Разом годин за модулем 1	**150**	**30**	**–**	**20**	**100**

5. Теми лабораторних занять

№ пор.	Назва теми	Кількість годин
		денна ф. н.
	Змістовний модуль 1. Базовий. Основи Data Science-проекту. Розвідувальний аналіз даних	
1	Лабораторна робота 1. Створення основи типового Data Science-проекту	2
2	Лабораторна робота 2. Маніпулювання даними	2
3	Лабораторна робота 3. Розвідувальний аналіз даних. Візуалізація	2
	Змістовий модуль 2. Просунутий рівень. Моделювання. Основні задачі та алгоритми	
4	Лабораторна робота 4. Побудова моделей класифікації	2
5	Лабораторна робота 5. Побудова моделей регресії	2
6	Лабораторна робота 6. Пошук асоціативних правил	2
7	Лабораторна робота 7. Побудова ARIMA-моделі часового ряду і прогнозування на її основі	2
8	Лабораторна робота 8. Веб-скрепінг	2
	Змістовий модуль 3. Засоби автоматизації розробки та розгортання Data Science-проекту	
9	Лабораторна робота 9. Створення дашбордів та розгортання Data Science-проекту	2
10	Захист фінального проєкта (РР)	2
	Разом годин	**20**

6. Самостійна робота

№ пор.	Назва теми	Кількість годин
		денна ф. н.
1	Тема 1. Введення у дисципліну.	5
2	Тема 2. Маніпулювання даними.	5
3	Тема 3. Розвідувальний аналіз даних. Візуалізація. Зниження розмірності	10
4	Тема 4. Моделювання. Класифікація.	10
5	Тема 5. Моделювання. Регресія.	10
6	Тема 6. Моделювання. Пошук асоціативних правил	5
7	Тема 7. Моделювання. Аналіз часових рядів. ARIMA-моделі часового ряду. Прогнозування	15
8	Тема 8. Веб-скрепінг	5
9	Тема 9. Створення дашбордів та розгортання Data Science-проекту	5
10	Фінальний проєкт (РР)	20
11	Підсумковий контроль: диф. залік	10
	Разом годин	**100**

7. Методи навчання

Пояснювально-ілюстративні, репродуктивні (опитування, тестування, розв’язування задач, виконання вправ за зразком).

Лекції, лабораторні роботи, консультації, самостійна робота.

Лекції викладаються з використанням мультимедійних засобів.

Самостійне опрацювання навчального матеріалу виконується з використанням конспекту лекцій, відеоматеріалів, основної та додаткової навчальної літератури, інформаційних ресурсів.

8. Методи контролю

Облік відвідування, опитування, захист лабораторних робіт, комплекти тестових завдань для проведення підсумкового контролю.

10. Розподіл балів, що отримують студенти

Критерії оцінювання		
Вид роботи	Зміст	Бали
Робота на лекціях	Активна участь у дискусіях, розгляд практичних кейсів	10
Завдання з ЛР (виконання, захист)	Лабораторна робота 1. Створення основи типового Data Science-проекту	5
	Лабораторна робота 2. Маніпулювання даними	5
	Лабораторна робота 3. Розвідувальний аналіз даних. Візуалізація	5
	Лабораторна робота 4. Побудова моделей класифікації	5
	Лабораторна робота 5. Побудова моделей регресії	5
	Лабораторна робота 6. Пошук асоціативних правил	5
	Лабораторна робота 7. Побудова ARIMA-моделі часового ряду і прогнозування на її основі	5
	Лабораторна робота 8. Веб-скрепінг	5
	Лабораторна робота 9. Створення дашбордів та розгортання Data Science-проекту	5
	Мініпроєкт	25
Контроль	Тестування за змістовим модулем 1	10
	Тестування за змістовим модулю 2	10
	Усього балів	**100**

Шкала оцінювання: національна та ECTS

Сума балів за 100-бальною шкалою	Оцінка в ECTS	Значення оцінки ECTS	Критерії оцінювання	Рівень компетентності	Оцінка за національною шкалою
					іспит, диференційо-ваний залік
90–100	A	відмінно	Студент виявляє особливі творчі здібності, вміє самостійно здобувати знання, без допомоги викладача знаходить та опрацьовує необхідну інформацію, вміє використовувати набуті знання і вміння для прийняття рішень у нестандартних ситуаціях, переконливо аргументує відповіді, самостійно розкриває власні обдарування і нахили	Високий (творчий)	відмінно
82–89	B	дуже добре	Студент вільно володіє вивченим обсягом матеріалу, застосовує його на практиці, вільно розв'язує вправи і задачі у стандартних ситуаціях, самостійно виправляє допущені помилки, кількість яких незначна	Достатній (конструктивно-варіативний)	добре
74–81	C	добре	Студент вміє зіставляти, узагальнювати, систематизувати інформацію під керівництвом викладача; вцілому самостійно застосовувати її на практиці; контролювати власну діяльність; виправляти помилки, серед яких є суттєві,добирати аргументи для підтвердження думок		
64–73	D	задовільно	Студент відтворює значну частину теоретичного матеріалу, виявляє знання і розуміння основних положень;з допомогою викладача може аналізувати навчальний матеріал, виправляти помилки, серед яких є значна кількість суттєвих	Середній (репродук-тивний)	задовільно
60–63	E	достатньо	Студент володіє навчальним матеріалом на рівні, вищому за початковий, значну частинуйого відтворює на репродуктивному рівні		
35–59	FX	незадовільно з можливістю повторного складання семестрового контролю	Студент володіє матеріалом нарівні окремих фрагментів, що становлять незначну частину навчального матеріалу	Низький (рецептивно-продуктивний)	незадовільно
1–34	F	незадовільно з обов'язковим повторним вивченням залікового кредиту	Студент володіє матеріалом нарівні елементарного розпізнання і відтворення окремих фактів, елементів, об'єктів		

10. Методичне забезпечення

1. Силабус з навчальної дисципліни «Інтелектуальний аналіз даних» для студентів денної форми навчання зі спеціальності 123 – «Комп'ютерна інженерія» освітньо-професійної програми «Комп'ютерна інженерія» освітнього ступеня «Магістр», 2023.

2. Електронний лабораторний практикум з навчальної дисципліни «Інтелектуальний аналіз даних» для студентів денної форми навчання зі спеціальності 123 – «Комп'ютерна інженерія» освітньо-професійної програми «Комп'ютерна інженерія» освітнього ступеня «Магістр», 2024. URL: https://vgamaley.github.io/DS-book-lab/

11. Рекомендована література

Основна

3. Casas, Pablo. 2018. *Data Science Live Book*. https://livebook.datascienceheroes.com/.

4. Garrett Grolemund, Hadley Wickham. 2018. *R for Data Science*. http://r4ds.had.co.nz/index.html.

5. Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.

6. Casas, Pablo. 2020. *funModeling: Exploratory Data Analysis and Data Preparation Tool-Box*. https://livebook.datascienceheroes.com.

7. Chan, Chung-hong, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker. 2018. *Rio: A Swiss-Army Knife for Data File i/o*.

8. Chan, Chung-hong, and Thomas J. Leeper. 2021. *Rio: A Swiss-Army Knife for Data i/o*. https://github.com/leeper/rio.

9. Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges.

2021. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

10. Conway, Joe, Dirk Eddelbuettel, Tomoaki Nishiyama, Sameer Kumar Prayaga, and Neil Tiffin. 2021. *RPostgreSQL: R Interface to the PostgreSQL Database System*. https://CRAN.R-project.org/package=RPostgreSQL.

11. Cooper, Nicholas. 2017. *Reader: Suite of Functions to Flexibly Read Data from Files*. https://CRAN.R-project.org/package=reader.

12. Garrett Grolemund, Hadley Wickham. 2018. *R for Data Science*. http://r4ds.had.co.nz/index.html.

13. Hadley Wickham, Lionel Henry, Romain Francois. 2018. "Introduction to Dplyr. Translation of Andrey Ogurtsov." Documentation. http://rpubs.com/aa989190f363e46d/dplyr_intro.

Hahsler, Michael. 2019. *arulesViz: Visualizing Association Rules and Frequent Itemsets*. https://CRAN.R-project.org/package=arulesViz.

14. ———. 2021. *arulesViz: Visualizing Association Rules and Frequent Itemsets*. https://github.com/mhahsler/arulesViz.

15. Hahsler, Michael, Christian Buchta, Bettina Gruen, and Kurt Hornik. 2022. *Arules: Mining Association Rules and Frequent Itemsets*. https://github.com/mhahsler/arules.

16. Hahsler, Michael, Bettina Gruen, and Kurt Hornik. 2005. "Arules – A Computational Environment for Mining Association Rules and Frequent Item Sets." *Journal of Statistical Software* 14 (15): 1–25. https://doi.org/10.18637/jss.v014.i15.

17. Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

Додаткова

1. Sydorenko, V., Perekrest, A., Shendryk, V., Shendryk, S. (2023). Machine Learning Optimization of Air Heating Time in the Heating Control System of a Smart House. In International Conference «New Technologies, Development and

Application». Lecture Notes in Networks and Systems, vol 707. Springer, Cham. pp. 36-44. https://doi.org/10.1007/978-3-031-34721-4_5

2. Guchenko M., Sydorenko V., Belska V., Liutenko M., Fesenko N. DComFra Project Learning Module M20 Advanced Spreadsheets in Mathematical Modeling Tasks of Electrical and Computer Engineers Education. Proceedings of the 20th IEEE International Conference on Modern Electrical and Energy Systems, MEES 2021.

3. Hester, Jim, and Hadley Wickham. 2021. *Odbc: Connect to ODBC Compatible Databases (Using the DBI Interface)*. https://CRAN.R-project.org/package=odbc.

4. Hyndman, Athanasopoulos, R. J. 2018. *Forecasting: Principles and Practice, 2nd Edition*. OTexts: Melbourne, Australia. https://otexts.com/fpp2/.

5. ———. 2021. *Forecasting: Principles and Practice, 3nd Edition*. OTexts: Melbourne, Australia.

6. Slabchenko Olesia, Siebert Xavier, Sydorenko Valeriy. 2016. "Development of Models for Imputation of Data from Social Networks on the Basis of an Extended Matrix of Attributes." *Eastern-European Journal of Enterprise Technologies* 4 (2-82): 24–34.

Інформаційні ресурси

7. wikipedia. 2018. "Cross-Industry Standard Process for Data Mining." Article. URL: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining. (дата звернення 26.01.2024)

8. Temple Lang, Duncan. 2022a. *RCurl: General Network (HTTP/FTP/...) Client Interface for r*. URL: https://CRAN.R-project.org/package=RCurl. (дата звернення 26.01.2024).

9. Posit. URL: https://posit.co/. (дата звернення 26.01.2024).

10. Quarto. URL: https://quarto.org/. (дата звернення 26.01.2024).

11. Shiny._URL: https://shiny.posit.co/. (дата звернення 26.01.2024).