КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ

Кафедра теорії та технології програмування

«ЗАТВЕРДЖУЮ»
Заступник декана
з навчальної роботи
_____ Олена КАШПУР
«07» травня 2021 р.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
АКТУАЛЬНІ ПРОБЛЕМИ «DATA MINING»

для студентів

галузь знань	**12 «Інформаційні технології»**
спеціальність	**122 «Комп'ютерні науки»**
освітній рівень	**магістр**
освітня програма	**«Інформатика»**
вид дисципліни	**обов'язкова**

Форма навчання	**денна**
Навчальний рік	**2021/2022**
Семестр	**2**
Кількість кредитів ECTS	**4**
Мова викладання, навчання та оцінювання	**українська**
Форма заключного контролю	**іспит**

Викладачі: **к.ф.-м.н., асистент Андрій Криволап**

Пролонговано: на 20__/20__ н.р. _____(_____) «__»___ 20__р.

на 20__/20__ н.р. _____(_____) «__»___ 20__р.

КИЇВ – 2021

Розробник: Андрій КРИВОЛАП, к.ф.-м.н., асистент кафедри теорії та технології програмування

ЗАТВЕРДЖЕНО

Завідувач кафедри
теорії та технології програмування
_____ Микола НІКІТЧЕНКО

Протокол № 10 від « 27 » квітня 2021 року

Схвалено гарантом освітньо-наукової програми «Інформатика»

_____ Степан ШКІЛЬНЯК

« 6 » травня 2021 року

Схвалено науково-методичною комісією факультету комп'ютерних наук та кібернетики

Протокол від « 6 » травня 2021 року № 10

Голова науково-методичної комісії _____ Людмила ОМЕЛЬЧУК

« 6 » травня 2021 року

1. Мета дисципліни – поглиблення знань з інтелектуального аналізу даних та штучного інтелекту, вивчення основних підходів до розв'язання основних задач – це задачі класифікації, кластеризації, пошуку асоціативних правил.

2. Попередні вимоги до опанування або вибору навчальної дисципліни

1. *Знати:* базові поняття штучного інтелекту та методів оптимізації; мати сучасні уявлення про основні задачі, що вирішуються в рамках штучного інтелекту та аналізу даних.
2. *Вміти:* описувати задачу аналізу даних, визначати атрибути та тип задачі, будувати модель.
3. *Володіти елементарними навичками:* з аналізу даних, методів оптимізації.

3. Анотація навчальної дисципліни:

Навчальна дисципліна «Актуальні проблеми Data Mining» є складовою освітньо-професійної програми підготовки фахівців за другим (*магістерським*) рівнем вищої освіти *галузі знань* 12 „Інформаційні технології" зі *спеціальності 122* „Комп'ютерні науки", *освітньо-професійної програми* –„Інформатика".

Дана дисципліна є обов'язковою навчальною дисципліною за ***програмою "Інформатика"***.

Викладається в 2 семестрі 1 курсу магістратури в обсязі 120 годин ***(4 кредити ECTS),*** зокрема: *лекції – 26 год., лабораторних занять – 12 год., самостійна робота – 80 год., консультації – 2 год.* У курсі передбачено **2** *частини* та 2 *контрольні роботи*. Завершується дисципліна – **іспитом в 1 семестрі.**

В результаті вивчення навчальної дисципліни студент повинен:

знати: основні задачі, що вирішує інтелектуальний аналіз даних; основні алгоритми, етапи та підходи, вирішення розглянутих задач; основні стандарти Data mining; методи оцінки основних характеристик побудованих моделей;

вміти: досліджувати запропонований набір даних; обирати найкращі методи для вирішення задач класифікації та кластеризації в залежності від вхідних параметрів та вимог; виконувати попередню обробку даних.

Для допуску до дисципліни „Актуальні проблеми Data Mining" освітньо-професійної програми «Інформатика» студент повинен опанувати компетентності та результати навчання, які надає дисципліна „Штучний інтелект" програми «Інформатика».

4. Завдання (навчальні цілі):

набуття знань, умінь та навичок (компетентностей) на рівні новітніх досягнень у програмуванні, відповідно освітньої кваліфікації «Магістр з комп'ютерних наук».

Зокрема:

- СК3. Здатність до дослідження та аналізу надвеликих масивів даних із складною неоднорідною і/або невизначеною структурою для прийняття зважених бізнес-рішень;
- СК8. Здатність вирішувати складні задачі інтелектуальної обробки даних з використанням еволюційного моделювання, нейромережних технологій, застосування обчислювального інтелекту для розв'язання практичних задач в різних галузях професійної діяльності.

5. Результати навчання за дисципліною

Результат навчання (1. знати; 2. вміти; 3. комунікація; 4. автономність та відповідальність)		Форми (та/або методи і технології) викладання і навчання	Методи оцінювання та пороговий критерій оцінювання (за необхідності)	Відсоток у підсумковій оцінці з дисципліни
Код	**Результат навчання**			
РН1.1	*Знати основні поняття, задачі та етапи інтелектуального аналізу даних.*	*Лекція, самостійна робота*	*Контрольна робота, іспит*	*15%*
РН1.2	*Знати основні методи інтелектуального аналізу даних, штучних нейронних мереж.*	*Лекція, самостійна робота*	*Контрольна робота, іспит*	*15%*
РН2.1	*Вміти аналізувати задачу та обирати адекватний метод аналізу даних, оцінювати точність застосованих методів, виділяти суттєві ознаки*	*Лекція, самостійна робота, лабораторні заняття*	*Контрольна робота, іспит, захист лабораторних робіт*	*50%*
РН3.1	*Обґрунтовувати власний погляд на задачу та спосіб її розв'язання, спілкуватися з колегами з питань застосування методів інтелектуального аналізу даних*	*Лекція, самостійна робота*	*Захист реферату, поточне оцінювання*	*10%*
РН4.1	*Організовувати свою самостійну роботу для досягнення результату*	*Лекція, самостійна робота, лабораторні заняття*	*Захист реферату, захист лабораторних робіт*	*10%*

6. Співвідношення результатів навчання дисципліни із програмними результатами навчання

Результати навчання дисципліни **Програмні результати навчання**	**РН 1.1**	**РН 1.2**	**РН 2.1**	**РН 3.1**	**РН 4.1**
(з опису освітньої програми)					
ПРН 3. Опанувати нові інструменти роботи з даними, здійснюючи обробку веб-логів, текст-аналіз і машинне навчання, для прогнозування бізнес-процесів та ситуаційного управління, сентимент-аналізу відгуків, розробки рекомендаційних систем для сфери електронної комерції, медіа, соціальних мереж, банкінгу, реклами тощо.		+	+	+	+
ПРН 4. Аналізувати великі дані та моделювати високорівневі абстракції у великих наборах даних різної природи, проектувати сховища великих даних, для видобутку даних і знань, візуалізовувати великі дані, будувати і оцінювати регресивні моделі, що генеруються на основі великих даних	+	+	+	+	

7. Схема формування оцінки

7.1 Форми оцінювання студентів:

- семестрове оцінювання:

1. Контрольні роботи: РН 1.1, РН 1.2, РН 2.1 – 2×10 = 20 балів / 12 балів

2. Лабораторні роботи: РН 2.1, РН 4.1– 15 балів / 9 балів

3. Реферат: РН 3.1, РН 4.1 – 15 балів / 9 балів

4. Поточне оцінювання (активна робота на заняттях): РН 2.1, РН 3.1, РН 4.1 – 5 балів

- підсумкове оцінювання:

- максимальна кількість балів які можуть бути отримані студентом: 40 балів;
- результати навчання які будуть оцінюватись: РН 1.1, РН 1.2, РН 2.1
- форма проведення: письмова форма.

Види завдань:

Структура екзаменаційної роботи та критерії оцінювання:

1. Теоретичне запитання (РН 1.1 – РН 1.2).
2. Теоретичне запитання (РН 1.1 – РН 1.2).
3. Задача (РН 2.1).
4. Задача (РН 2.1).

Критерії оцінювання екзаменаційної роботи

Завдання	Вид завдання	Максимальний бал (відсоток)	Всього балів (відсотків)
Завдання 1	Теоретичне запитання	8 балів (20 %)	8 балів (20 %)
Завдання 2	Теоретичне запитання	8 балів (20 %)	8 балів (20 %)
Завдання 3	Задача	12 балів (30 %)	12 балів (30 %)
Завдання 4	Задача	12 балів (30 %)	12 балів (30 %)
Всього			**40 балів (100%)**

Студент допускається до екзамену якщо семестрі набрав не менше ніж 30 балів та отримав не менше мінімальної порогової кількості балів за лабораторні та контрольні роботи.

Для отримання загальної позитивної оцінки з дисципліни оцінка за іспит має бути не менше 24 балів.

Питання на іспит

1. Поняття даних. Набір даних і їх атрибутів.
2. Вимірювання. Шкали.
3. Типи наборів даних. Формати зберігання даних.
4. Поняття метаданих.
5. Задачі та методи Data Mining.
6. Класифікація та властивості методів Data Mining.

7. Задача класифікації.
8. Точність класифікації: оцінка рівня помилок.
9. Алгоритм побудови елементарних правил (1-rule).
10. Алгоритми класифікації. Наївний баєсівський класифікатор.
11. Застосування нейронних мереж для задач класифікації.
12. Методи побудови дерев прийняття рішень.
13. Алгоритм найближчого сусіда.
14. Постановка задачі пошуку асоціативних правил. Алгоритм Apriory та його різновиди.
15. Постановка задачі кластеризації, загальна схема кластеризації.
16. Ієрархічні алгоритми кластеризації, алгоритм k-means та метод найближчого сусіда.
17. Застосування нейронних мереж для задач кластеризації (Карта Кохонена).
18. Адаптивні методи кластеризації.

7.2 Організація оцінювання:

Терміни проведення форм оцінювання:

1. Контрольна робота 1 : до 10 тижня семестру.
2. Контрольна робота 2: до 19 тижня семестру.
3. Захист реферату: до 19 тижня семестру.
4. Лабораторні роботи: протягом семестру.
5. Поточне оцінювання: протягом семестру.

Студент має право на одне перескладання контрольної роботи із можливістю отримання за роботу максимально 8 балів. Термін перескладання визначається викладачем.

За відсутності студента з поважних причин перескладання КР здійснюється відповідно до «Положення про організацію освітнього процесу».

7.3 Шкала відповідності оцінок

Відмінно / Excellent	90-100
Добре / Good	75-89
Задовільно / Satisfactory	60-74
Незадовільно / Fail	0-59

8. Структура навчальної дисципліни.
Тематичний план лекцій та лабораторних занять

№ лекції	Назва лекції	Кількість годин		
		Лекції	Лабораторні заняття	Самостійна робота
Частина 1. **Основні поняття Data Mining. Задача класифікації даних**				
1	**Тема 1.** Поняття даних. Задачі Data Mining. Стандарти Data Mining. Методи Data Mining **Самостійна робота:** Модель MapReduce. Класифікація методів DM. Властивості методів DM.	2		6
2	**Тема 2.** Постановка задачі класифікації. Точність класифікації. Оцінка рівня помилок **Самостійна робота:** Оцінка рівня помилок за допомогою крос перевірки. Питання вибору тестової множини.	2		6
3	**Тема 3.** Алгоритм побудови елементарних правил (1-rule). Наївний баєсівський класифікатор **Самостійна робота:** Обмеження наївного баєсівського класифікатору. Задача визначення спаму. **Лабораторна робота:** Застосування алгоритмів побудови елементарних правил та наївного баєсівського класифікатору для задачі класифікації даних.	2	2	8
4	**Тема 4.** Методи побудови дерев прийняття рішень. Алгоритм найближчого сусіда **Самостійна робота:** Проблема перенавчання для дерев прийняття рішень та способи її вирішення. Області застосувань дерев рішень. Основні проблеми алгоритму найближчого сусіда і шляхи їх вирішення. **Лабораторна робота:** Застосування алгоритмів побудови дерева прийняття рішень та алгоритму найближчого сусіда для задачі класифікації даних.	4	2	10
5	**Тема 5.** Застосування нейронних мереж для задач класифікації **Самостійна робота:** Подання вхідних даних для штучних нейронних мереж. Вибір архітектури мережі.	2	1	8
	Контрольна робота		1	
	Всього за частиною 1	12	6	38
Частина 2. **Задача пошуку асоціативних правил. Задача кластеризації даних**				
6	**Тема 6.** Задача пошуку асоціативних правил. Алгоритм Apriori та його різновиди **Самостійна робота:** Узагальнені асоціативні правила. Модифікації алгоритму Apriori. Алгоритм FP-Growth. **Лабораторна робота:** Застосування алгоритму Apriori для пошуку асоціативних правил.	2	2	8
7	**Тема 7.** Задачі кластеризації. Виділення характеристик. Визначення метрики. Приклади метрик **Самостійна робота:** Стратегії кластеризації. Багатовимірні евклідові простори та «прокляття вимірності».	2		6
8	**Тема 8.** Ієрархічні алгоритми кластеризації. Дендрограми. Алгоритм Single-link. Алгоритм Complete-link.	2	2	6

	Самостійна робота: Ефективність ієрархічної кластеризації. Ієрархічна кластеризація у неевклідових просторах. **Лабораторна робота:** Застосування ієрархічних алгоритмів для задачі кластеризації.			
8	**Тема 9.** Неієрархічні алгоритми кластеризації. Алгоритм k-means. Алгоритм найближчого сусіда. **Самостійна робота:** Алгоритм fuzzy k-means. Алгоритм Бредлі, Файяда та Рейна. Алгоритм CURE. **Лабораторна робота:** Застосування алгоритмів k-means та найближчого сусіда для задачі кластеризації.	2	2	8
9	**Тема 10.** Самоорганізаційна Карта Кохонена. Алгоритм навчання. Відображення кластерів. **Самостійна робота:** Початкова ініціалізація карти. Питання вибору конфігурації сітки.	2		8
10	**Тема 11.** Адаптивні методи кластеризації. Визначення якості кластеризації. Показники чіткості **Самостійна робота:** Нечіткі алгоритми кластеризації.	3		6
	Контрольна робота	1		
	Всього за частиною 2	14	6	42
	ВСЬОГО	**26**	**12**	**80**

Загальний обсяг *120 год.,* в тому числі:
Лекцій **– *26** год.*
Лабораторні заняття **– *12** год.*
Консультації – ***2** год.*
Самостійна робота **- *80** год.*

9. Рекомендовані джерела

Основні:

1. Барсегян и др. Методы и модели анализа данных: OLAP и DM. – СПб., 2004
2. Berry, Michael J. A. “DM techniques: for marketing, sales, and customer relationship management “/ Michael J.A. Berry, Gordon Linoff. – 2nd ed.
3. Larose, Daniel T. “Discovering knowledge in data: an introduction to DM” / Daniel T. Larose
4. Leskovec J. Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman // Stanford Univ. – 2010.
5. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
6. Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.
7. R. Agrawal, R. Srikant. "Fast Discovery of Association Rules", In Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994.

Додаткові:

8. G. Lee,U. Yun A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Generation Computational Systems 68:89–110 p., 2017.
9. S. Rustogi, M. Sharma, S. Morwal Improved Parallel Apriori Algorithm for Multi-cores. IJ Inf Technol Comput Sci 4:18–23p., 2017.
10. M.K. Gupta, P. Chandra A comparative study of clustering algorithms. In: Proceedings of the 13th INDIACom-2019; IEEE Conference ID: 461816; 6th International Conference on “Computing for Sustainable Global Development”, 2019.
11. К. Шеннон. Работы по теории информации и кибернетике. М. Иностранная литература, 1963
12. W. Buntine. A theory of classification rules. 1992.
13. Добыча данных в сверхбольших базах данных / В. Ганти, Й. Герке, Р. Рамакришнан // Открытые системы, №9-10, 1999.
14. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
15. R. M. Hristev, "Artifical Neural Networks"
16. R. Srikant, R. Agrawal. "Mining Generalized Association Rules", In Proc. of the 21th International Conference on VLDB, Zurich, Switzerland, 1995.
17. J.S. Park, M.-S. Chen, and S.Y. Philip, "An Effective HashBased Algorithm for Mining Association Rules", In Proc. ACM SIGMOD Int’l Conf. Management of Data, ACM Press, New York, 1995.
18. S. Brin et al., "Dynamic Itemset Counting and Implication Rules for Market Basket Data", In Proc. ACM SIGMOD Int’l Conf. Management of Data, ACM Press, New York, 1997.