Міністерство освіти і науки України

Харківський національний університет імені В.Н. Каразіна

Кафедра теоретичної та прикладної системотехніки

“ЗАТВЕРДЖУЮ”

Проректор з науково-педагогічної роботи

_____ Олександр ГОЛОВКО

“____” _____2022 р.

Робоча програма навчальної дисципліни

Теоретичні основи методології Big Data processing

рівень вищої освіти другий (магістерський)

спеціальність 123 «Комп’ютерна інженерія»

освітня програма Комп’ютерна інженерія

вид дисципліни за вибором

факультет комп’ютерних наук

2022 / 2023 навчальний рік

Програму обговорено та рекомендовано до затвердження вченою радою факультету комп'ютерних наук
«28» червня 2022 року, протокол №10

РОЗРОБНИКИ ПРОГРАМИ:
кандидат технічних наук, доцент кафедри теоретичної та прикладної системотехніки **Стрілець Вікторія Євгенівна**

Програму схвалено на засіданні кафедри теоретичної та прикладної системотехніки
Протокол від «11» червня 2022 року, №12

Завідувач кафедри теоретичної та прикладної системотехніки
_____ Сергій ШМАТКОВ.

Програму погоджено з гарантом освітньої програми «Комп'ютерна інженерія»

Гарант освітньої програми «Комп'ютерна інженерія»
_____ Олена ТОЛСТОЛУЗЬКА

Програму погоджено методичною комісією факультету комп'ютерних наук
Протокол від «24» червня 2022 року № 9

Голова методичної комісії факультету комп'ютерних наук
_____ Анатолій БЕРДНІКОВ

ВСТУП

Програма навчальної дисципліни «Теоретичні основи методології Big Data processing» розроблена відповідно до освітньо-професійної програми підготовки другого (магістерського) рівня спеціальності 123 «Комп'ютерна інженерія» освітньої програми «Комп'ютерна інженерія».

1. Опис навчальної дисципліни

1.1. **Метою** викладання навчальної дисципліни є:

засвоєння студентами основ наукових і математичних положень, моделей і методів, що лежать в основі обробки даних і технології машинного навчання для дослідження складних систем та процесів, зокрема аналізу і удосконалення складних комп'ютерних системам; вироблення навичок з використання алгоритмів машинного навчання та програмних засобів, призначених для розв'язання задач обробки даних.

1.2**. Основними завданнями вивчення навчальної дисципліни є** вивчення і набуття навичок застосування:

- методів попередньої обробки даних, отриманих під час проведення експериментів або вивчення процесів різної природи;.
- методів розв'язання задачі класифікації даних;
- методів ідентифікації математичних моделей систем та процесів;
- методів оцінювання інформативності (значущості) змінних;
- методів прогнозування даних.

В ході вивчення дисципліни у студента повинні формуватися такі компетентності.

Загальні компетентності (ЗК)

ЗК01. Вміння виявляти, ставити та вирішувати проблеми за професійним спрямуванням.

ЗК02. Здатність проведення досліджень на відповідному рівні.

ЗК04. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

ЗК05. Здатність до творчого, креативного і абстрактного мислення, аналізу та синтезу.

ЗК06.Здатність приймати обґрунтовані рішення.

ЗК07. Здатність розробляти проекти і управляти ними.

ЗК08. Здатність оцінювати та забезпечувати якість виконуваних робіт.

ЗК09. Здатність працювати як індивідуально, так і в команді.

Спеціальні (фахові, предметні) компетентності (ФК)

ФК01 Здатність обґрунтовано обирати та застосовувати фундаментальні знання і моделі, а також технології створення та використання прикладного і спеціалізованого програмного забезпечення для розв'язування складних професійних задач і проблем комп'ютерної інженерії.

ФК03. Здатність до дослідження, системного аналізу та забезпечення безперервності бізнес/операційних процесів, концепцій, теорій, принципів і методів нових технологій, включаючи технології розумних, мобільних, зелених і безпечних обчислень.

ФК04. Здатність застосовувати системний підхід до вирішення науково-технічних завдань комп'ютерної інженерії.

ФК05. Здатність досліджувати, розв'язувати складні професійні завдання і проблеми на основі розуміння технічних аспектів забезпечення контролю якості продукції.

ФК06. Здатність досліджувати, розробляти та впроваджувати засоби і системи автоматизації проектування до розробки компонентів комп'ютерних систем та мереж, Інтернет додатків, кіберфізичних систем тощо.

ФК07. Здатність застосовувати комплексний підхід до вирішення експериментальних завдань модернізації та реконструкції комп'ютерних систем та мереж, різноманітних вбудованих і розподілених додатків, зокрема з метою підвищення їх ефективності.

ФК10. Здатність проводити та організовувати, планувати науково-дослідницьку діяльність в сфері комп'ютерної інженерії, відповідно вітчизняним та світовим стандартам і вимогам.

ФК11. Здатність аргументувати вибір методів розв'язування складних спеціалізованих задач і проблем, критично оцінювати отримані результати, обґрунтовувати та захищати прийняті рішення.

ФК12. Здатність створювати дослідницькі групи для проведення аналізу та обробки великих масивів даних.

ФК13. Здатність перетворювати формальні моделі в напрямку отримання практично необхідної комп'ютерної моделі та ставити задачі збереження і обробки даних.

ФК14. Здатність здійснювати наукові та/або прикладні дослідження у галузі комп'ютерної інженерії із застосуванням сучасних експериментальних і теоретичних методів моделювання процесів, критично оцінювати результати досліджень та інновацій, презентувати результати досліджень та формувати науково-технічну звітність.

1.3. **Кількість кредитів** – 5

1.4. **Загальна кількість годин** – 150

1.5. **Характеристика навчальної дисципліни**	
За вибором	
Денна форма навчання	Заочна (дистанційна) форма навчання
Рік підготовки	
1-й	1-й
Семестр	
1-й	1-й
Лекції	
32 год.	год.
Практичні, семінарські заняття	
16 год.	год.
Лабораторні заняття	
0 год.	год.
Самостійна робота	
102 год.	год.
Індивідуальні завдання	
-год.	

1.6. Відповідно до вимог освітньо-кваліфікаційного рівня підготовки за результатами вивчення дисципліни студенти повинні –

знати:

- основні види машинного навчання;
- задачі, які відносяться до навчання без вчителя;
- задачі, які відносяться до навчання з вчителем
- методи машинного навчання для розв'язання задач класифікації, регресії та кластеризації даних;

уміти:

- здійснювати вибір методів машинного навчання для розв'язання задач класифікації, регресії та кластеризації даних;

- проводити верифікацію методів машинного навчання та оцінку їх якості застосування на основі існуючих критеріїв;
- розв'язувати задачі попереднього аналізу даних, класифікації, регресії та кластеризації даних із застосуванням спеціалізованих бібліотек та мов програмування;
- застосовувати підходи і методи машинного навчання для створення ефективних систем автоматизації складних процесів;
- пояснювати, кількісно та якісно оцінювати, корегувати отримані результати;

придбати навички:

- статистичної обробки даних;
- застосування методів машинного навчання для розв'язання задач класифікації, регресії та кластеризації даних;
- проведення верифікації методів, оцінки якості математичних методів на основі існуючих критеріїв;
- вирішення задач машинного навчання з застосуванням спеціалізованих бібліотек;

мати уявлення:

- про роль методів машинного навчання у створенні сучасних складних технічних систем; перспективах розвитку методів машинного навчання; про основні проблеми розробки сучасного програмного забезпечення для розв'язання задач аналізу, інтелектуальної обробки даних та ін.

В результаті вивчення дисципліни у студента повинні формуватися такі *програмні результати навчання (ПРН)*.

ПРН01. Знати і розуміти сучасні методи наукових досліджень, організації та планування експерименту, збирання даних та моделювання в комп'ютерних системах.

ПРН02. Знати і розуміти наукові і математичні положення, що лежать в основі функціонування програмних і програмно-технічних комп'ютерних засобів, систем та мереж, Інтернету речей, систем для оброблення великих даних.

ПРН04. Знати і розуміти принципи системного аналізу та забезпечення безперервності бізнес/операційних процесів, концепцій, теорій, принципів і методів нових технологій, включаючи технології розумних, мобільних, зелених і безпечних обчислень.

ПРН06. Мати фундаментальні знання і розуміння моделей, а також технологій створення та використання прикладного і спеціалізованого програмного забезпечення розв'язування професійних задач і проблем комп'ютерної інженерії.

ПРН07. Знати засоби автоматизації проектування до розробки компонентів комп'ютерних систем та мереж, Інтернет додатків, кіберфізичних систем тощо.

ПРН10. Вміти формулювати та розв'язувати задачі у галузі комп'ютерної інженерії, що пов'язані з процедурами спостереження об'єктів, вимірювання, контролю, діагностування і прогнозування з урахуванням загальнолюдських цінностей, суспільних, державних та виробничих інтересів.

ПРН11. Мати навички автономного і самостійного навчання у сфері комп'ютерної інженерії і дотичних галузей знань, аналізувати власні освітні потреби та об'єктивно оцінювати результати навчання.

ПРН15. Мати навички планування та виконання експериментальних і теоретичних досліджень та випробувань, вибору для цього придатних методи та інструменти, здійснювання статистичної обробки даних, оцінки адекватності отриманих результатів.

ПРН16. Вміти досліджувати, розробляти та впроваджувати засоби і системи автоматизації проектування до розробки компонентів комп'ютерних систем та мереж, Інтернет додатків, кіберфізичних систем тощо.

ПРН17. Застосовувати, інтегрувати, розробляти, впроваджувати та удосконалювати сучасні інформаційні технології, науково-технічні методи і моделі, фізичні та математичні фундаментальні знання в галузі комп'ютерної інженерії.

ПРН18. Здатність аргументувати вибір методів розв’язування складних спеціалізованих задач і проблем, критично оцінювати отримані результати, обґрунтовувати та захищати прийняті рішення.

ПРН20. Вільно користуватися державною та іноземною мовами, усно і письмово для представлення і обговорення результатів досліджень та інновацій, забезпечення бізнес\операційних процесів та питань професійної діяльності в галузі комп’ютерної інженерії.

ПРН21. Зрозуміло і недвозначно доносити власні висновки з проблем комп’ютерної інженерії, а також знання та пояснення, що їх обґрунтовують.

ПРН23. Здатність адаптуватись до нових ситуацій, обґрунтовувати, приймати та реалізовувати у межах компетенції рішення.

ПРН24. Усвідомлювати необхідність навчання впродовж усього життя з метою поглиблення набутих та здобуття нових фахових знань, удосконалення креативного мислення.

ПРН25. Якісно виконувати роботу та досягати поставленої мети з дотриманням вимог професійної етики як самостійно, так і в команді.

ПРН27. Здатність володіти науково-методичними знаннями в галузі комп’ютерної інженерії; формулювати ідеї, концепції з метою використання в роботі освітнього та наукового спрямування.

ПРН28. Виявляти зв’язки між сучасними концепціями в організації освітнього процесу та наукового пізнання в області комп’ютерної інженерії.

2. Тематичний план навчальної дисципліни

Тема 1. Машинне навчання як напрям штучного інтелекту.

Етапи розвитку штучного інтелекту, виникнення і виокремлення машинного навчання. Визначення машинного навчання, його основні напрямки (з вчителем, без вчителя, з підкріпленням). Формальне поняття «навчання».

Тема 2. Задачі машинного навчання.

Задачі навчання з вчителем. Задачі навчання без вчителя. Способи навчання та оцінка його якості. Перенавчання і недонавчання, узагальнююча здатність.

Тема 3. Попередня обробка даних.

Статистичний аналіз даних. Виявлення та обробка відсутніх значень. Кодування нечислових ознак. Масштабування і стандартизація. Кодування.

Тема 4. Задача класифікації та методи її розв’язання.

Загальна постановка задачі класифікації. Наївний байєсівський класифікатор. Дерева рішень. Випадкові ліси. Логістична регресія. K-найближчих сусідів. Метод опорних векторів.

Тема 5. Задача регресії та методи її розв’язання.

Загальна постановка задачі регресії. Базова регресійна модель. Лінійна регресія. Методи нелінійної регресії (дерева рішень, випадкові ліси, метод опорних векторів).

Тема 6. Задача кластеризації та методи її розв’язання.

Загальна постановка задачі кластеризації. Метод k-середніх. Метод c-середніх. Агломеративна (ієрархічна) кластеризація.

3. Структура навчальної дисципліни

Назви розділів і тем	Кількість годин					
	Денна форма					
	Всього	у тому числі:				
		Л	ПЗ	Лаб. роб.	Інд.	СР
1	2	3	4	5	6	7
Тема 1. Машинне навчання як напрям штучного інтелекту.	14	2	2			10
Тема 2. Задачі машинного навчання.	25	6	2			17

Тема 3. Попередня обробка даних.	25	6	2			17
Тема 4. Задача класифікації та методи її розв'язання.	31	6	6			21
Тема 5. Задача регресії та методи її розв'язання.	25	6	2			17
Тема 6. Задача кластеризації та методи її розв'язання.	30	6	2			20
Усього годин	**150**	**32**	**16**			**102**

4. Теми практичних занять

№ п/п	Назва теми	Кількість годин
1	Програмні засоби машинного навчання. Мова програмування Python. Бібліотеки numpy, pandas.	2
2	Пошук та завантаження наборів даних для аналізу. Статистичний аналіз.	2
3	Попередня обробка даних. Робота з відсутніми даними.	2
4	Задача класифікації. Наївний байєсівський класифікатор.	2
5	Задача класифікації. Дерева рішень і випадкові ліси.	2
6	Задача класифікації. K-найближчих сусідів. Метод опорних векторів.	2
7	Задача регресії. Лінійна регресія. Методи нелінійної регресії.	2
8	Задача кластеризації. Метод k-середніх. Ієрархічна кластеризація.	2
	Разом	16

5. Завдання для самостійної роботи

№ п/п	Зміст	Кількість годин
1	Ознайомитися з мовою програмування Python, бібліотеками машинного навчання, вбудованими набори даних.	17
2	Розглянути існуючі підходи до опрацювання відсутніх даних у наборах даних.	17
3	Розглянути метрики оцінювання якості роботи методів машинного навчання.	21
4	Провести порівняльний аналіз методів класифікації даних.	17
5	Провести порівняльний аналіз методів регресії.	20
6	Підготовка до підсумкової контрольної роботи	10
	Разом	102

6. Індивідуальні завдання

7. Методи контролю

Контроль роботи студентів при вивченні дисципліни і засвоєння ними навчального матеріалу здійснюється на практичному зайнятті шляхом проведення контрольних опитувань і захисту звітів з практичних завдань. Підсумковий контроль здійснюється при виконанні 1 контрольної роботи і письмової залікової відповіді.

Студенти, що не виконали впродовж семестру 1 контрольну роботу, а також не представили і не захистили звіти з практичних завдань, до заліку не допускаються.

Залік проводиться у вигляді тестової роботи або письмової відповіді на теоретичне і практичне питання.

При дистанційному навчанні видача практичних завдань та контроль їх виконання здійснюється за допомогою сервісу дистанційного навчання Google Classroom. Лекційні заняття проводяться із використанням сервісу відео-конференцій Google Meet.

8. Схема нарахування балів

Бали за поточний контроль знань впродовж семестру (по темах)						Разом	Залік	Сума
					Контрольна робота, передбачена навчальним планом			
Т1, 2	Т3	Т4	Т5	Т6	1			
8	8	16	8	8	12	60	40	100

Т1, Т2 ... – теми розділів.
За темами Т1, 2 студент отримує 8 балів за виконання практичної роботи 1.
За темою Т3 студент отримує 8 балів за виконання практичної роботи 2.
За темою Т4 студент отримує 16 балів за виконання практичних робіт 3 і 4.
За темою Т5 студент отримує 8 балів за виконання практичної роботи 5.
За темою Т6 студент отримує 8 балів за виконання практичної роботи 6.

Критерії оцінювання знань студентів за практичні роботи

Вимоги	Кількість балів
.▪ Завдання відзначається повнотою виконання без допомоги викладача. ▪ Визначає рівень поінформованості, потрібний для прийняття рішень. Вибирає інформаційні джерела,. ▪ Робить висновки і приймає рішення у ситуації невизначеності. Володіє уміннями творчо-пошукової діяльності.	8
▪ Завдання – повні, з деякими огріхами, виконані без допомоги викладача. ▪ Планує інформаційний пошук; володіє способами систематизації інформації; ▪ Робить висновки і приймає рішення у ситуації невизначеності. Володіє уміннями творчо-пошукової діяльності.	6-7
▪ Завдання відзначається неповнотою виконання без допомоги викладача. ▪ Студент може зіставити, узагальнити, систематизувати інформацію під керівництвом викладача; вільно застосовує вивчений матеріал у стандартних ситуаціях.	4-5
▪ Завдання відзначається неповнотою виконання за консультацією викладача. ▪ Застосовує запропонований вчителем спосіб отримання інформації, має фрагментарні навички в роботі з підручником, науковими джерелами; ▪ Вибирає відомі способи дій для виконання фахових методичних завдань.	3
Завдання відзначається фрагментарністю виконання за консультацією викладача або під його керівництвом.	1-2

Критерії оцінювання знань студентів за контрольну роботу

Вимоги	Кількість балів
Повнота виконання завдання повна, студент здатен формулювати закони та закономірності, структурувати судження, умовиводи, доводи, описи.	9-12
Повнота виконання завдання повна, студент здатен формулювати операції, правила, алгоритми, правила визначення понять.	5-8
Повнота виконання завдання елементарна, студент здатен вибирати відомі способи дій для виконання фахових завдань.	3-5
Повнота виконання завдання фрагментарна.	1-2

Критерії оцінювання залікових робіт студентів

Вимоги	Кількість балів
Показані всебічні систематичні знання та розуміння навчального матеріалу; безпомилково виконані завдання.	35-40
Показані повні знання навчального матеріалу; помилки, якщо вони є, не носять принципового характеру.	30-35
Показано повне знання необхідного навчального матеріалу, але допущені помилки.	20-30
Показано повне знання необхідного навчального матеріалу, але допущені суттєві помилки	10-20
Показано недосконале знання навчального матеріалу, допущені суттєві помилки.	5-10
Показано недосконале знання навчального матеріалу, допущені суттєві помилки, які носять принциповий характер; обсяг знань не дозволяє засвоїти предмет.	1-5

Шкала оцінювання

Сума балів за всі види навчальної діяльності протягом семестру	Оцінка	
	для чотирирівневої шкали оцінювання	для дворівневої шкали оцінювання
90 – 100	відмінно	зараховано
70-89	добре	
50-69	задовільно	
1-49	незадовільно	не зараховано

9. Рекомендована література

Основна література

1. MacKay David J.C. Information Theory, Inference and Learning Algorithms / David J.C. MacKay. – Cambridge University Press, 2003. – 628 p.
2. Bishop Christopher M. Pattern Recognition and Machine Learning / Christopher M. Bishop. – New York: Springer, 2006. – 738 p.
3. Harrison M. Machine Learning Pocket Reference / Mett Harrison. – O'Reilly Media, Inc., 2019. – 320 p.
4. Müller A. C. Introduction to Machine Learning with Python / Andreas C. Müller, Sarah Guido. – O'Reilly Media, Inc., 2016. – 393 p.
5. VanderPlas J. Python Data Science HandBook. Essential tools for working with data / Jake VanderPlas. – O'Reilly, 2018. – 576 p.

6. Andrew Ng. Machine Learning Yearning. – 2018. – 118 p.
7. Rasmussen C.E. Gaussian Processes for Machine Learning / C.E. Rasmussen, C.K.I. Williams. – Cambridge, Massachusetts: MIT Press, 2006. – 248 p.
8. Merphy K. Machine Learning: A Probabilistic Perspective / Kevin Merphy. (https://mitpress.mit.edu/books/machine-learning-0).

Допоміжна

1. Довбиш А.С. Основи теорії розпізнавання образів: навч. посіб / А.С. Довбиш, І.В. Шелехов. – Суми: Сумський державний університет, 2015. – 109 с.
2. Coelho L. P. Building machine learning systems with Python / Luis Pedro Coelho, Willi Richert. – Packt Publishing Ltd., 2015. – 302 p.
3. Raschka S. Python machine Learning: Machine Leraning and Deep Learning with Python, scikit-learn, and TensorFlow2 / Sebastian Raschka, Vahid Mirjalili. – Packt Publishing, 2019. – 772 p.
4. Cielen D. Intoduction Data Science: Big Data, Machine Learning, and more, using Python tools / Davy Cielen, Arno D.B. Meysman, Mohamed Ali. – Manning, 2016. – 320 p.
5. Fenner M. Machine Learning with Python for Everyone / Mark Fenner. – O'Reilly Media, Inc., 2019.
6. Sutton Richard S., Barto Andrew G. Reinforcement Learning: An Introduction. – Cambridge, Massachusetts: MIT Press, 2018. – 426 p.
7. Hastie Trevor, Tibshirani Robert, Friedman Jerome. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. – New York: Springer, 2009. – 745 p.

10. Посилання на інформаційні ресурси в Інтернеті, відео-лекції, інше методичне забезпечення

1. Machine Learning Course. URL: https://www.youtube.com/watch?v=jGwO_UgTS7I&list=PLoROMvodv4rMiGQp3WXShtMGgzqpfVfbU.
2. Datasets. URL: https://www.kaggle.com, https://archive.ics.uci.edu/ml/datasets.php