МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики
Кафедра кібербезпеки

Затверджено
На засіданні кафедри кібербезпеки
факультету прикладної математики та інформатики
Львівського національного університету імені Івана Франка
(Протокол №9/24 від 29 серпня 2024 р.)

Завідувач кафедри .

Петро ВЕНГЕРСЬКИЙ

Силабус з навчальної дисципліни
“Інтелектуальний аналіз великих даних”,
що викладається в межах ОПП Технології штучного інтелекту в кібербезпеці
другого (магістерського) рівня вищої освіти для здобувачів з спеціальності 125 – Кібербезпека та захист інформації

Львів 2024 р.

Назва дисципліни	Інтелектуальний аналіз великих даних
Адреса викладання дисципліни	Львівський національний університет імені Івана Франка, вул. Університетська 1, м. Львів, Україна, 79000
Факультет та кафедра, за якою закріплена дисципліна	Факультет прикладної математики та інформатики Кафедра кібербезпеки
Галузь знань, шифр та назва спеціальності	12 – інформаційні технології 125 – кібербезпека та захист інформації
Викладачі дисципліни	Гірна Олександра Йосипівна, кандидат фіз.-мат. наук, доцент кафедри кібербезпеки (лекції та лабораторні заняття)
Контактна інформація викладачів	oleksandra.hirna@lnu.edu.ua; Головний корпус ЛНУ ім. І. Франка, каб. 380. м. Львів, вул. Університетська, 1
Консультації з питань навчання по дисципліні відбуваються	Консультації проводять раз на тиждень згідно з оприлюдненим розкладом консультацій викладача. Можливі онлайн консультації через Zoom чи Microsoft Teams. Для погодження часу онлайн консультацій слід писати на електронну пошту викладача.
Сторінка курсу	https://e-learning.lnu.edu.ua/course/view.php?id=6132
Інформація про дисципліну	Дисципліна “Інтелектуальний аналіз великих даних” є вибірковою дисципліною з спеціальності 125 – кібербезпека та захист інформації для освітньої програми Технології штучного інтелекту в кібербезпеці, яка викладається у 1-му семестрі другого (магістерського) рівня освіти в обсязі 4.5 кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Курс спрямований на формування у студентів розуміння базових концепцій інтелектуального аналізу даних та алгоритмів штучного інтелекту з розбудови розумних моделей, а також їх реалізацій засобами мови програмування Python.
Мета та цілі дисципліни	Метою курсу є формування у студентів системи знань базових концепцій та алгоритмів інтелектуального аналізу даних і практичних навичок їх застосування для побудови та аналізу моделей на основі даних.
Література для вивчення дисципліни	**Основна література:** 1. Харченко В. О. Основи машинного навчання : навч. посіб. / В. О. Харченко. – Суми : СДУ, 2023. – 264 с. 2. Шарадкін Д.М., Субач І.Ю., Микитюк А.В. Інструментальні засоби Python для моделювання та системного аналізу часових рядів при вирішенні задач кіберзахисту інформаційно-комунікаційних систем: навч. пос. / Шарадкін Д.М., Субач І.Ю., Микитюк А.В.; ІСЗЗІ КПІ ім. Ігоря Сікорського. – Київ : КПІ ім. Ігоря Сікорського, 2023. - 139 с. 3. Інтелектуальний аналіз даних та машинне навчання. Частина Базові методи та засоби аналізу даних / Я. В. Іванчук, В. І. Месюра, А. А. Яровий, О. Д. Манжілевський – Вінниця : ВНТУ, 2021. – 69 с. 4. Susan E. McGregor Practical Python Data Wrangling and Data Quality / O’Reilly Media, 2022. – 578 p. 5. Bernd Klein Machine Learning with Python Tutorial / Bodenceo, 2021. – 453 p. **Додаткова література:**

	6. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. - К.: КПІ ім. Ігоря Сікорського», 2018. - 300 с. 7. Черняк О.І.. Інтелектуальний аналіз даних: Підручник / О.І. Черняк, П.В. Захарченко ; Київ. нац. ун-т ім. Т. Шевченка. — К. : Знання, 2014. - 599 с. 8. Гороховатський В.О., Творошенко І.С. Методи інтелектуального аналізу та оброблення даних: навч. посібник. – Харків: ХНУРЕ, 2021. - 92 с. 9. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. — К.: ІСЗЗІ КПІ ім. Ігоря Сікорського», 2018. - 300 с. 10. Ситник В. Ф. Інтелектуальний аналіз даних (дейтамайнінг): Навч. По-сібник/ В. Ф. Ситник, М.Т. Краснюк - К: КНЕУ, 2007. - 376 с. **Інформаційні ресурси в Інтернет** 11. Сайт ЛНУ ім.Івана Франка http://www.mmf.lnu.edu.ua/ar/1739 12. Національний інститут стандартів і технологій (NIST), Стандарти науки про дані та аналізу великих даних: https://bigdatawg.nist.gov/standards/ 13. Група спеціальних інтересів з виявлення знань та інтелектуального аналізу даних (SIGKDD) Асоціації обчислювальної техніки (Association for Computing Machinery, ACM): https://www.kdd.org/ 14. Процес міжгалузевого стандарту для інтелектуального аналізу даних (CRISP-DM): https://www.datascience-pm.com/crisp-dm-2/ 15. Data Mining Group (DMG): https://www.dmg.org/ Стандарти мови розмітки предиктивних моделей (PMML): http://dmg.org/pmml/v4-3/GeneralStructure.html *Рекомендовані онлайн курси* 1. https://www.coursera.org/specializations/data-science-python 2. https://www.coursera.org/learn/python-for-applied-data-science-ai 3. https://www.coursera.org/specializations/big-data 4. https://www.coursera.org/learn/python-project-for-data-engineering
Обсяг курсу	Загальний обсяг: 135 годин. Аудиторних занять: 64 год., з них 32 год. лекцій та 32 год. лабораторних робіт. Самостійної роботи: 71 год.
Очікувані результати навчання	У результаті вивчення навчальної дисципліни студент має набути таких компетентностей: **знати:** • основні концепції маніпуляції з даними та формування масивів даних; • методи машинного навчання аналізу даних; • методи графічної аналітики великих даних; • методи моделювання, прогнозування та керування на базі гібридних систем; • базові алгоритми класичного навчання з учителем: та без учителя машинного навчання; • сутність базових алгоритмів побудови ансамблів; • методи регресійного, кластерного та асоціативного аналізу великих даних; • методи та прийоми перевірки, навчання та покращення якості моделі. **вміти:** • застосовувати методи: ✓ класифікації, групування, очищення та візуалізації даних, у тому числі враховувати особливості роботи з часовими рядами; ✓ дерев рішень;

	✓ регресійних моделей; ✓ вибору та інженерії ознак моделі; ✓ зменшення розмірності моделі; ✓ кластеризації; ✓ асоціативних правил; ✓ побудови ансамблів; • обирати алгоритм навчання системи штучного інтелекту та реалізувати його засобами мови програмування Python для вирішення конкретної практичної задачі.
Ключові слова	Інтелектуальний аналіз даних, машинне навчання, навчання з учителем, навчання без учителя, криві навчання, великі дані.
Формат курсу	Очний. Проведення лекцій, лабораторних робіт і консультацій.
Теми	Теми подані у Схемі курсу нижче
Підсумковий контроль, форма	Залік у кінці семестру
Пререквізити	Для вивчення курсу студенти потребують базові знання з таких дисциплін: 1) Моделі та методи дискретної математики; 2) Основи математичного аналізу та їх застосування; 3) Обчислювальна геометрія та алгебра; 4) Програмування; 6) Застосування теорії ймовірностей в кібербезпеці
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Презентації, лекції Модульний контроль Індивідуальні завдання
Необхідне обладнання	Лабораторія з обладнаними робочими станціями, з'єднаними в комп'ютерну мережу. Python, Jupyter Notebook.
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться за 100-бальною шкалою. 75 балів нараховують за виконання 5 лабораторних завдань та 25 балів – за оволодіння теоретичним матеріалом курсу (модульний контроль - 25 балів) **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, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в письмовій роботі студента є підставою для її незарахування викладачем, незалежно від масштабів плагіату чи обману. **Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та лабораторні заняття курсу. Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх видів письмових робіт та індивідуальних завдань, передбачених курсом. **Література.** Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі

третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих.

Політика виставлення балів. Враховуються бали набрані при поточному тестуванні, самостійній роботі та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час лабораторного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін. Жодні форми порушення академічної доброчесності не толеруються.

Критерії оцінювання знань студентів	**Бали рейтингу**	**Макс. к-сть балів**
1. Бали поточної успішності за виконання індивідуальних завдань		
Критерії оцінювання (5*15 бали)	**75 балів**	
Студент в повному обсязі володіє навчальним матеріалом, вільно самостійно та аргументовано його викладає під час захисту індивідуальних завдань, глибоко та всебічно розкриває зміст теоретичних питань. Реалізоване програмне забезпечення пройшло перевірку на плагіат та повністю виконує умову завдання.	**15**	
Студент достатньо повно володіє навчальним матеріалом, обґрунтовано його викладає під час захисту індивідуальних завдань, в основному розкриває зміст теоретичних питань. Але при викладанні деяких питань не вистачає достатньої глибини та аргументації. Реалізоване програмне забезпечення містить окремі несуттєві неточності та незначні помилки.	**14-7**	
Студент не в повному обсязі володіє навчальним матеріалом. Фрагментарно, поверхнево (без аргументації та обґрунтування) викладає його під час захисту лабораторного завдання, недостатньо розкриває зміст теоретичних питань, допускаючи при цьому суттєві неточності, програмна реалізація індивідуального завдання частково виконана.	**6-1**	
Студент не виконав індивідуальне завдання та не володіє матеріалом.	**0**	
2. Модульний контроль		
Критерії оцінювання (25 балів)	**25**	
Модульний контроль проводиться наприкінці семестру. Модуль містить **25 тестових питань.**		
***Критерії оцінювання вирішення тестів (25*1 бали)*:**		
Відповідь вірна	**1**	
Відповідь невірна	**0**	
Загальна кількість балів по завершенні вивчення дисципліни	**100**	

Додаткові бали / або зарахування певних тем можна отримати за результатами **неформального та/або інформального навчання за** тематикою даної дисципліни. Визнання та зарахування результатів такого навчання відбувається у відповідності до наданих документів про неформальне та/або інформальне навчання.

	Жодні форми порушення академічної доброчесності не толеруються.
Питання до контролю	1. Засоби та стандартизовані процеси аналізу даних 2. Основні етапи CRISP-метолології 3. Описовий аналіз даних з Pandas 4. Особливості графічної аналітики великих даних 5. Виявлення та обробка відсутніх значень, дублікатів, викидів 6. Групування, таблиці спряженості 7. Шкалювання даних 8. Описові статистики для числових та категорійних ознак 9. Візуалізація даних з Seaborn, Matplotlib, Plotly 10. Розподіли ознак, гістограми, коробкові діаграми, точкові графіки 11. Особливості обробки часових рядів. Модулі time, datetime 12. Основні складові, види та задачі машинного навчання 13. Масштабування алгоритмів машинного навчання 14. Дерева рішень. Критерії якості: ентропія, невизначеність Джині. 15. Критерій зупинення алгоритму. Відсікання гілок. Вилучення правил 16. Лінійна регресія. Метод найменших квадратів. Теорема Гауса-Маркова 17. Декомпозиція дисперсії із зміщенням. Регуляризація моделі 18. Метод максимуму правдоподібності 19. Лінійна класифікація. Логістична регресія 20. L_2-регуляризація логістичної моделі 21. Зменшення помилок моделі та поняття ансамблю. Теорема Кондорсе 22. Ідеї методів бутстрепу та бегінгу 23. Алгоритм випадкового лісу. Параметри підвищення точності моделі 24. Метод k-найближчих сусідів Визначення класу нового об'єкта. Вибір параметра k. 25. Значущість та вибір ознак моделі 26. Лассо-регресія і рідж-регресія 27. L_1-регуляризація 28. Метод головних компонент. Зменшення розмірності даних 29. Методи кластеризації 30. Метод k-середніх. Покрокова реалізація. Вибір кількості кластерів
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

Схема курсу

Тиж.	Тема, план, короткі тези	Форма діяльності (заняття)	Література	Завдання, год.	Термін виконання
1-2	**Тема 1.** Засоби та стандартизовані процеси аналізу великих даних. Можливості Python Групування, таблиці спряженості Шкалювання даних. Дослідження даних за допомогою зведеної статистики	лекція, самостійна робота	[1-10]	4 10	2 тижні
		лаб.	[1-10]	4	
3	**Тема 2.** Візуалізація даних з Seaborn, Matplotlib, Plotly. Дослідження даних за допомогою графіків	лекція, самостійна робота		2 5	1 тиждень
		лаб.	[1-10]	2	

4	**Тема 3.** Складові, види та задачі машинного навчання. Техніки очищення даних. Виявлення та обробка відсутніх значень, дублікатів, викидів. Особливості обробки часових рядів	лекція, самостійна робота	[1-10]	2 5	1 тиждень
		лаб.	[1-10]	2	
5	**Тема 4.** Навчання з учителем. Класифікація: Дерева рішень. Метод k-найближчих сусідів	лекція, самостійна робота	[1-10]	2 5	1 тиждень
		лаб.	[1-10]	2	
6-7	**Тема 5.** Дерева рішень. Критерії якості та зупинення алгоритму. Відсікання гілок. Вилучення правил	лекція, самостійна робота	[1-10]	4 10	2 тижні
		лаб.	[1-10]	4	
8	**Тема 6.** Ентропія, невизначеність Джині. Ключові параметри дерева	лекція, самостійна робота	[1-10]	2 5	1 тиждень
		лаб.	[1-10]	2	
9-10	**Тема 7.** Лінійна регресія. Метод найменших квадратів. Теорема Гауса-Маркова. Декомпозиція дисперсії із зміщенням. Перевірка та криві навчання. Регуляризація моделі. Метод максимуму правдоподібності Лінійна класифікація. Логістична регресія. L_2-регуляризація логістичної моделі	лекція, самостійна робота	[1-10]	4 7	2 тижні
		лаб.	[1-10]	4	
11	**Тема 8.** Зменшення помилок моделі та поняття ансамблю. Теорема Кондорсе. Ідеї методів бутстрепу та бегінгу	лекція, самостійна робота	[1-10]	2 4	1 тиждень
		лаб.	[1-10]	2	
12	**Тема 9.** Алгоритм випадкового лісу. Параметри підвищення точності моделі	лекція, самостійна робота	[1-10]	2 4	1 тиждень
		лаб.	[1-10]	2	
13	**Тема 10.** Значущість ознак. Вибір ознак моделі. Лассо-регресія і рідж-регресія. L_1-регуляризація	лекція, самостійна робота	[1-10]	2 4	1 тиждень
		лаб.	[1-10]	2	
14	**Тема 11.** Навчання без учителя. Метод головних компонент. Зменшення розмірності даних Методи кластеризації. Метод k-середніх. Покрокова реалізація. Вибір кількості кластерів	лекція, самостійна робота	[1-10]	2 4	1 тиждень
		лаб.	[1-10]	2	
15	**Тема 12.** Пошук асоціативних правил. Алгоритм Apriori	лекція, самостійна	[1-10]	2 4	1 тиждень

		робота			
		лаб.	[1-10]	2	
16	**Тема 13.** Аналіз часових рядів. Методи прогнозування. Згладжування. Лаги часових рядів. Стаціонарність, одиничний корінь	лекція, самостійна робота	[1-10]	2 4	1 тиждень
		лаб.	[1-10]	2	