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 № 15/23 від 29 серпня 2023 р.)

Завідувач кафедри ____ П.С.Венгерський

Силабус з навчальної дисципліни "Методи інтелектуального аналізу даних", що викладається в межах ОПП Кібербезпека першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 125 – кібербезпека та захист інформації

Львів 2023 р.

Назва дисципліни	Методи інтелектуального аналізу даних
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1
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, канд. фіз.-мат. наук, доцент, доцент кафедри вищої математики
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	Консультація проводиться за розкладом консультацій викладача. Можливі дистанційні консультації за попередньою домовленістю.
Сторінка курсу	https://e-learning.lnu.edu.ua/course/view.php?id=6132
Інформація про дисципліну	Дисципліна “ Методи інтелектуального аналізу даних ” є вибірковою дисципліною для спеціальності 125 – кібербезпека та захист інформації для освітньої програми Кібербезпека, яка викладається в 7-му семестрі в обсязі 5-ох кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено так, щоб надати учасникам розуміння базових концепцій інтелектуального аналізу даних та алгоритмів розділу штучного інтелекту – машинного навчання, що займається розбудовою розумних моделей, а також їх реалізацій засобами мови програмування Python. У дисципліні представлені алгоритми та інструментальні засоби первинної статистичної обробки та очищення даних, різноманітні засоби групування та візуалізації; алгоритми класичного навчання з вчителем та без вчителя, а також ансамблеві методи покращення якості моделі.
Мета та цілі дисципліни	**Метою** навчальної дисципліни 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 вибіркових даних. **Завданням** вивчення навчальної дисципліни є сформувати у студентів теоретичні знання та практичні навички в області обробки та представлення даних; методів класифікації, регресії, кластеризації, пошуку асоціативних правил, побудови ансамблів.
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. Посібник/ 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/ 16. Стандарти мови розмітки предиктивних моделей (PMML): http://dmg.org/pmml/v4-3/GeneralStructure.html
Обсяг курсу	Загальний обсяг: 150 годин. Аудиторних занять: 64 год., з них 32 годин лекцій та 32 години лабораторних занять. Самостійної роботи: 86 годин.
Очікувані результати навчання	**У результаті вивчення навчальної дисципліни студент повинен *знати:*** • основні концепції маніпуляції з даними та формування масивів даних; • основні поняття та методи машинного навчання; • методи моделювання, прогнозування та керування на базі гібридних систем; • обчислювального інтелекту; • сутність базових алгоритмів класичного навчання з учителем: та без учителя машинного навчання; • сутність базових алгоритмів побудови ансамблів; • 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. **Курс забезпечує набуття таких компетентностей: ІК, КЗ 1, КЗ 5, КФ 2, КФ 5, КФ 12 та програмних результатів навчання: ПРН 2-3, ПРН 4, ПРН 6, ПРН 10-13, ПРН 16, ПРН 28-30.**
Ключові слова	Інтелектуальний аналіз даних, машинне навчання, навчання з учителем, навчання без учителя, криві навчання.
Формат курсу	Очний, дистанційний
Теми	Завдання та стандарти інтелектуального аналізу даних Дослідницький аналіз даних та задачі описової статистики Основні поняття методів машинного навчання Навчання з учителем: Дерева рішень. Метод k-найближчих сусідів. Регресія. Логістична регресія. Метод максимуму правдоподібності Перевірка моделі та криві навчання Навчання без учителя: Метод головних компонент. Кластеризація. Метод k-середніх. Пошук асоціативних правил Ансамблеві методи Аналіз часових рядів
Підсумковий контроль, форма	Залік у кінці семестру
Пререквізити	Для вивчення дисципліни студенти потребують базові знання з курсів “Моделі та методи дискретної математики”, “Основи математичного аналізу та їх застосування”, “Прикладні основи обчислювальної геометрії та алгебри”, “Застосування теорії ймовірностей в кібербезпеці”, “Програмування”.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Презентації, лекції, лабораторні роботи, індивідуальні завдання, індивідуальні доповіді, самостійна робота. Лекції та лабораторні: інформаційно-рецептивний метод, репродуктивний метод, евристичний метод, метод проблемного викладу. Самостійна робота: репродуктивний метод, дослідницький метод.
Необхідне обладнання	Комп’ютер, Internet.
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням: • лабораторні роботи: 50% семестрової оцінки; максимальна кількість балів – 50; • реферат і доповідь: 20% семестрової оцінки; максимальна кількість балів – 20; • контрольний тест: по 30% семестрової оцінки; максимальна кількість балів – 30. **Академічна доброчесність**: Очікується, що роботи студентів будуть їх самостійними дослідженнями чи міркуваннями. Виявлення ознак академічної недоброчесності в письмовій роботі студента є підставою для її незарахування викладачем, незалежно від масштабів плагіату чи обману.

	Відвідування занять є важливою складовою навчання. Усі студенти зобов'язані відвідувати усі лекції та лабораторні заняття курсу, дотримуватися термінів виконання індивідуальних завдань. **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 опитуванні, виконанні самостійних робіт, бали тестувань. Враховуються активність студентів під час занять, своєчасність виконання завдань, не допускається списування та плагіат. 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. Виявлення та обробка відсутніх значень, дублікатів, викидів 5. Групування, таблиці спряженості 6. Шкалювання даних 7. Описові статистики для числових та категорійних ознак 8. Візуалізація даних з Seaborn, Matplotlib, Plotly 9. Розподіли ознак, гістограми, коробкові діаграми, точкові графіки 10. Особливості обробки часових рядів. Модулі time, datetime 11. Основні складові, види та задачі машинного навчання 12. Дерева рішень. Критерії якості: ентропія, невизначеність Джині. 13. Критерій зупинення алгоритму. Відсікання гілок. Вилучення правил 14. Лінійна регресія. Метод найменших квадратів. Теорема Гауса-Маркова 15. Декомпозиція дисперсії із зміщенням. Регуляризація моделі 16. Метод максимуму правдоподібності 17. Лінійна класифікація. Логістична регресія 18. L_2-регуляризація логістичної моделі 19. Зменшення помилок моделі та поняття ансамблю. Теорема Кондорсе 20. Ідеї методів бутстрепу та бегінгу 21. Алгоритм випадкового лісу. Параметри підвищення точності моделі 22. Метод k-найближчих сусідів Визначення класу нового об'єкта. Вибір параметра k. 23. Значущість та вибір ознак моделі 24. Лассо-регресія і рідж-регресія 25. L_1-регуляризація 26. Метод головних компонент. Зменшення розмірності даних 27. Методи кластеризації 28. Метод k-середніх. Покрокова реалізація. Вибір кількості кластерів
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенні курсу.

Схема курсу

Тиж.	Тема заняття	Форма діяльності (заняття)	Література	Заняття, год.	Самост. робота, год.
1.	Засоби та стандартизовані процеси аналізу даних. Можливості Python	лекція	[1-10]	2	5
	Описовий аналіз даних з Pandas Групування, таблиці спряженості	лабораторна		2	

2.	Шкалювання даних. Описові статистики для числових та категорійних ознак	лекція	[1-10]	2	5
		лабораторна		2	
3.	Візуалізація даних з Seaborn, Matplotlib, Plotly	лекція	[1-10]	2	6
	Розподіли ознак, гістограми, коробкові діаграми, точкові графіки	лабораторна		2	
4.	Основні складові, види та задачі машинного навчання	лекція	[1-10]	2	5
	Виявлення та обробка відсутніх значень, дублікатів, викидів. Особливості обробки часових рядів	лабораторна		2	
5.	Навчання з учителем. Класифікація: Дерева рішень. Метод k-найближчих сусідів	лекція	[1-10]	2	5
	Дерева рішень. Критерії якості та зупинення алгоритму. Відсікання гілок. Вилучення правил	лабораторна		2	
6.	Ентропія, невизначеність Джині. Ключові параметри дерева	лекція	[1-10]	2	5
	Захист ЛР №1	лабораторна		2	
7.	Лінійна регресія. Метод найменших квадратів. Теорема Гауса-Маркова	лекція	[1-10]	2	6
		лабораторна		2	
8.	Декомпозиція дисперсії із зміщенням. Перевірка та криві навчання. Регуляризація моделі. Метод максимуму правдоподібності	лекція	[1-10]	2	6
		лабораторна		2	
9.	Лінійна класифікація. Логістична регресія. L_2-регуляризація логістичної моделі	лекція	[1-10]	2	5
		лабораторна		2	
10.	Зменшення помилок моделі та поняття ансамблю. Теорема Кондорсе. Ідеї методів бутстрепу та бегінгу	лекція	[1-10]	2	6
	Захист ЛР №2	лабораторна		2	
11.	Алгоритм випадкового лісу. Параметри підвищення точності моделі	лекція	[1-10]	2	5
		лабораторна		2	
12.	Значущість ознак. Вибір ознак моделі. Лассо-регресія і рідж-регресія. L_1-регуляризація	лекція	[1-10]	2	6
		лабораторна		2	
13.	Навчання без учителя. Метод головних компонент. Зменшення розмірності даних	лекція	[1-10]	2	5
		лабораторна		2	
14.	Методи кластеризації. Метод k-середніх. Покрокова реалізація. Вибір кількості кластерів	лекція	[1-10]	2	5
		лабораторна		2	
15.	Пошук асоціативних правил. Алгоритм Apriori	лекція	[1-10]	2	5
	Доповідь за рефератом	лабораторна		2	
16.	Аналіз часових рядів. Методи прогнозування. Згладжування. Лаги часових рядів. Стаціонарність, одиничний корінь	лекція	[1-10]	2	6
	Тест	лабораторна		2	
	Всього 150 год.			**64**	**86**