МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК

КАФЕДРА ІНТЕЛЕКТУАЛЬНИХ ІНФОРМАЦІЙНИХ СИСТЕМ

"ЗАТВЕРДЖУЮ"
Перший проректор
Котляр Ю.В.
"30" серпня 2024 року

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

«ІНТЕЛЕКТУАЛЬНІ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ»

Спеціальність: 122 Комп'ютерні науки

Розробники	Калініна І.О.
	Гожий О.П.
Завідувач кафедри	Кондратенко Ю. П
Гарант освітньої програми	Гожий О.П.
Декан факультету	Бойко А.П.
Начальник НМВ	Шкірчак С. І.

Миколаїв – 2024 рік

1. Опис навчальної дисципліни

Найменування показника	Характеристика дисципліни	
Найменування дисципліни	Інтелектуальні технології аналізу та попередньої обробки даних	
Шифр курсу в освітній програмі	**ПП 2**	
Галузь знань	F – Інформаційні технології	
Спеціальність	F3 – Комп'ютерні науки	
Спеціалізація (якщо є)	-	
Освітня програма	Інтелектуальні інформаційні системи	
Рівень вищої освіти	Магістр	
Статус дисципліни	Нормативна	
Курс навчання	I	
Навчальний рік	2024-2025	
Номер семестрів:	Денна форма	Заочна форма
	1 семестр	-
Загальна кількість кредитів ЄКТС/годин	4 кредитів / 120 годин	
Структура курсу:	Денна форма	Заочна форма
– лекції	15	-
– практичні	30	-
– годин самостійної роботи студентів	75	-
Відсоток аудиторного навантаження	38%	-
Мова викладання	Українська	
Оригінальність навчальної дисципліни	**Авторський курс**	
Форма проміжного контролю	Тестування	
Форма підсумкового контролю	Залік	

2. Мета, завдання та результати вивчення дисципліни

Аналіз та попередня обробка даних – це процес перетворення необроблених даних у значущі за допомогою різних методів.

Дані в реальному світі "неочищені, сирі". Це означає, що вони можуть бути неповні, містити відсутні, непотрібні, повторювані дані, дані з шумом тощо. Використання такого типу необроблених даних у моделях машинного навчання не забезпечує її хорошої продуктивності. Щоб отримати хорошу та точну продуктивність, необхідно попередньо обробити дані, тобто перетворити необроблені дані на значущі. Метою попередньої обробки даних є підвищення точності та ефективності подальшого аналізу та моделювання.

Метою освоєння дисципліни «Інтелектуальні технології аналізу та попередньої обробки даних» є підготовка фахівців, здатних розв'язувати комплексні проблеми в галузі дослідницької діяльності у сфері аналізу даних для прийняття оптимальних рішень та використання результатів аналізу даних для уточнення наукових висновків та формування прогнозів щодо прийнятих рішень, що передбачає глибоке розуміння підходів до створення традиційних та створення нових поведінкових моделей.

Завдання:

1. Ознайомлення з основними методами та технологіями інтелектуального аналізу та попередньої обробки даних.
2. Навчання збору та підготовки даних для аналізу.
3. Розуміння та вміння застосовувати різноманітні алгоритми машинного навчання та інші методи інтелектуального аналізу даних для попередньої обробки даних.

4. Вивчення способів візуалізації та інтерпретації результатів аналізу даних.

5. Розвиток навичок створення моделей прогнозування на основі інтелектуального аналізу даних та їх застосування в різних сферах діяльності.

Пре реквізити:

Відповідно до освітньої програми необхідно попередньо оволодіти знаннями з дисциплін: «Вища математика», «Основи програмування», «Об'єктно-орієнтовне програмування», «Теорія ймовірностей та математична статистика», «Алгоритми та структури даних», «Системний аналіз», «Методи та системи штучного інтелекту».

З дисципліни «Вища математика» використовуються знання і навички з обчислення похідних і визначення інтегралів, рішення диференційних рівнянь.

З дисципліни «Основи програмування та алгоритмічні мови» використовуються знання і навички з розробки програм різноманітних структур.

З дисципліни «Основи дискретної математики» використовуються знання з детермінованих та імовірнісних автоматів.

З дисципліни «Теорія ймовірностей і математична статистика» використовуються знання стосовно ймовірності, випадкової величини, випадкової функції, випадкового процесу, функції розподілу, щільності ймовірності, статистичних моментів та їх оцінок.

З дисципліни «Алгоритми та методи обчислень» використовуються знання та навички з чисельного інтегрування, з чисельного рішення диференціальних рівнянь, інтерполяції та апроксимації функцій.

По реквізити:

Компетентності, знання та вміння, отримані в рамках вивчення даної дисципліни, можуть бути застосовані для отримання обґрунтованих результатів досліджень та підвищення наукового рівня кваліфікаційної роботи

Вивчення дисципліни дає *підґрунтя* для подальшого засвоєння нових та більш спеціалізованих програмних продуктів з ймовірнісно-статистичних методів моделювання та прогнозування тощо.

Під час вивчення дисципліни передбачається систематична практична робота студентів, як при виконанні спеціально підготовлених навчальних вправ, так і за реальними виробничими завданнями.

Очікувані результати навчання:

В результаті вивчення дисципліни студент

***має знати*:**

- концептуальні та методологічні знання у сфері аналізу даних для прийняття оптимальних рішень, а також дослідницькі навички, достатні для проведення наукових і прикладних досліджень в галузі аналізу даних на рівні останніх світових досягнень з комп'ютерної інженерії, IT-інфраструктур, інформаційних технологій;
- сучасні методи проведення досліджень у сфері прикладного аналізу даних для прийняття рішень в різних галузях науки шляхом вивчення теоретичних та практичних положень побудови поведінкових моделей функціональних залежностей та використання результатів аналізу даних для уточнення оптимальних параметрів прийняття рішень;
- наукові і математичні положення, що лежать в основі методів аналізу даних для отримання оптимальних параметрів прийняття рішень, методів побудови та дослідження математичних моделей та технологій адаптивних та інтелектуальних обчислень при аналізі даних.

***має вміти*:**

- ефективно здійснювати пошук та критичний аналіз інформації з різних джерел щодо методів та технологій аналізу даних;
- розв'язувати задачі синтезу та аналізу об'єктів дослідження при аналізі даних для прийняття оптимальних рішень;

- створювати та реалізовувати поведінкові математичні моделі опису функціональних залежностей при аналізі даних;
- застосовувати прикладні методи аналізу даних технології адаптивних та інтелектуальних обчислень для передбачення майбутніх станів досліджуваних об'єктів та процесів.
- застосовувати прикладні бібліотеки та програмні системи, які використовуються при машинному аналізі даних;
- володіти методами та технологіями програмування з використанням прикладних бібліотек та програмних систем, призначених для машинного аналізу реальних наборів даних.

Знання, які студенти набудуть при вивченні курсу «Інтелектуальні технології аналізу та попередньої обробки даних» будуть необхідними при подальшому навчанні та освоєнні фахових та спеціальних дисциплін, а також у виробничій діяльності з фахової спеціальності.

В результаті вивчення дисципліни студент отримує (*згідно зі стандартом вищої освіти України для другого (магістерського) рівня від 28.04.2022 № 393*):

Загальні компетентності:

– ЗК02. Здатність застосовувати знання у практичних ситуаціях.

Спеціальні (фахові, предметні) компетентності:

– СК04. Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проєктних рішень.
– СК06. Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі комп'ютерних наук.
– СК12. Здатність розробляти елементи програмного забезпечення для реалізації інтелектуальних технологій, а саме нейронних мереж, методів нечіткої логіки, еволюційних методів обчислень, біоінспірованих методів в інтелектуальних системах.
– СК13. Здатність розробляти та реалізовувати проєкти з інтелектуальних систем в тому числі за допомогою методів машинного і глибокого навчання, імітаційних моделей, застосовуючи нові інструментальні засоби розробки, сучасні бібліотеки мов програмування та сучасні візуальні технології.

Програмні результати навчання:

– РН 8. Розробляти математичні моделі та методи аналізу даних (включно з великим).
– РН 9. Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).
– РН 20. Розробляти та застосовувати елементи програмного забезпечення для реалізації інтелектуальних технологій, а саме нейронних мереж, методів нечіткої логіки, еволюційних методів обчислень, біоінспірованих методів в інтелектуальних системах.
– РН 21. Розробляти та застосовувати елементи програмного забезпечення для реалізації інтелектуальних технологій, а саме нейронних мереж, методів нечіткої логіки, еволюційних методів обчислень, біоінспірованих методів в інтелектуальних системах.

3. Програма навчальної дисципліни

Денна форма:

№	Теми	Лекції	Практичні	Самостійна робота
1	**Тема 1.** Основні поняття інтелектуальних технології аналізу та попередньої обробки даних. Порівняння методів машинного навчання та інтелектуального аналізу даних.	2	4	9
2	**Тема 2.** Методи ідентифікації та заповнення пропусків у структурованих наборах даних та часових рядах.	2	4	9
3	**Тема 3.** Методи і алгоритми виявлення, обробки аномальних значень та структурних змін в даних.	2	4	9
4	**Тема 4.** Методи та підходи вибору ознак при розв'язанні задач машинного навчання.	2	4	9
5	**Тема 5.** Системний підхід до нормалізації та стандартизації даних у задачах машинного навчання.	2	4	9
6	*Тестове опитування*	1	-	6
7	**Тема 6.** Методи та алгоритми згладжування даних. Системне використання методів фільтрації даних у задачах машинного навчання.	2	4	9
8	**Тема 7.** Інформаційна технологія ймовірнісне-статистичного аналізу та попередньої обробки даних.	2	4	9
9	*Підсумкове тестове опитування*	-	2	6
	Всього за курсом	**15**	**30**	**75**

4. Зміст навчальної дисципліни

4.1. План лекцій

№	Тема заняття / план
1	**Тема 1: Основні поняття інтелектуальних технології аналізу та попередньої обробки даних. Порівняння методів машинного навчання та інтелектуального аналізу даних.** **Зміст:** Постановка задачі аналізу даних. Сфери застосування задач прикладного аналізу даних. Поняття та мета попередньої обробки даних. Елементи структурованих даних. Розробка базової таблиці аналітики. Мета дослідження даних. Побудова звіту про якість даних. Виявлення проблем з якістю даних. Класифікація проблем машинного навчання. Методологія CRISP-DM. Аналіз його елементів. Детальний приклад використання методології CRISP-DM.
2	**Тема 2: Методи ідентифікації та заповнення пропусків у структурованих наборах даних та часових рядах.** **Зміст:** Загальний алгоритм обробки пропусків даних. Приклади використання методів ідентифікації та заповнення пропущених значень у структурованих наборах даних та часових рядах.

3	**Тема 3: Методи і алгоритми виявлення, обробки аномальних значень та структурних змін в даних.** **Зміст:** Загальний алгоритм методу ідентифікації та обробки викидів у масиві даних та структурних змін. Приклади використання методів обробки аномальних значень у наборах даних. Приклади використання методів виявлення структурних змін в часових рядах.
4	**Тема 4: Методи та підходи вибору ознак при розв'язанні задач машинного навчання.** **Зміст:** Розгляд задач генерації ознак. Методи вибору функцій (методи фільтрації, методи обгортки та група вбудованих методів), зворотне виключення, прямий вибір, виключення та вичерпний вибір функцій. Виявлення та видалення «непотрібних» предикторів. Розгорнутий приклад.
5	**Тема 5: Системний підхід до нормалізації та стандартизації даних у задачах машинного навчання.** **Зміст:** Загальний алгоритм нормалізації даних. Лінійні та нелінійні методи нормалізації даних. Приклади використання методів нормалізації на різних наборах даних.
6	**Тема 6: Методи та алгоритми згладжування даних. Системне використання методів фільтрації даних у задачах машинного навчання.** **Зміст:** Методи ковзного середнього, експоненційного згладжування, фільтра Савицького-Голея, фільтрації Калмана для вирішення задачі згладжування даних. Приклади використання методів і алгоритмів згладжування на різних наборах даних. Цифрова фільтрація. Оптимальна фільтрація. Імовірнісна фільтрація. Комплексне використання фільтрів різних типів для попередньої обробки даних.
7	**Тема 7: Інформаційна технологія ймовірнісне-статистичного аналізу та попередньої обробки даних.** **Зміст:** Побудова інформаційної технології ймовірнісне-статистичного аналізу та попередньої обробки даних. Системний підхід до аналізу та попередньої обробки даних. Побудова інформаційно-аналітичної системи аналізу та попередньої обробки даних.

4.2. План практичних занять

№	Тема заняття / план
1-2	**Тема 1.** Основні поняття інтелектуальних технології аналізу та попередньої обробки даних. Порівняння методів машинного навчання та інтелектуального аналізу даних. *Робота №1: Попередня обробка часових рядів.* **Завдання:** Виконати загальний статистичний аналіз часового ряду за варіантом. Перевірте часової ряд за допомогою статистичних тестів на нелінійність і нестаціонарність. Виявіть типи нелінійності та нестаціонарності часового ряду. Виконайте декомпозицію ряду на складові компоненти. Зробіть загальні висновки.
3-4	**Тема 2.** Методи ідентифікації та заповнення пропусків у структурованих наборах даних та часових рядах. *Робота №2: Виявлення відсутніх значень в часових рядах.* **Завдання:** Проаналізувати структуру набору даних. Проаналізувати та обробити пропущені спостереження часового ряду. Усунути перепустки шляхом заповнення їх попередніми значеннями ряду (метод LOCF). Візуалізувати результати перетворень. Виконати агрегування спостережень. Виконати декомпозицію часового ряду (метод STL). Проаналізувати зв'язок між значеннями часового ряду (наявність автокореляції). Виконати аналіз ACF та PACF. Виконати тести на автокореляцію (Дарбіна-Уотсона та Бройша-Готфрі). Виконати тести на перевірку стаціонарності. Виконати необхідну трансформацію часового ряду.

5	**Тема 3.** Методи і алгоритми виявлення, обробки аномальних значень та структурних змін в даних. *Робота №3: Виявлення аномалій та структурних змін в часових рядах.* **Завдання:** Виявити незвичайні спостереження в часовому ряду, для цього: виконати декомпозицію часового ряду на окремі складові (сезонну, тренд та залишки); застосувати до залишків один метод виявлення аномалій; відновити вихідний часовий ряд, паралельно обчислюючи верхню та нижню межі діапазону, до якого входять "нормальні" спостереження; уявіть результати графічно.
6	**Тема 3.** Методи і алгоритми виявлення, обробки аномальних значень та структурних змін в даних. *Робота №3: Виявлення аномалій та структурних змін в часових рядах.* **Завдання:** Візуалізувати вихідний часовий ряд та виявлені в ньому незвичайні спостереження. Дослідити вплив параметрів (окремо кожен параметр) функцій, що дозволяють виявити незвичайні спостереження в часовому ряду, шляхом ручного налаштування значень. Застосуйте метод «*E-Divisive with Medians*» (EDM) для виявлення структурних змін у часовому ряді.
7-8	**Тема 4.** Методи та підходи вибору ознак при розв'язанні задач машинного навчання. *Робота №4: Методи відбору множини змінних (ознак).* **Завдання:** Виконати вибір оптимального підмножини змінних і знайти модель з оптимальним числом предикторов. Відібрати підмножини змінних моделі методами покрокового включення та виключення змінних. Виконати методи перевірочної вибірки та оцінки оптимальної множини змінних. Видібрати оптимальну модель з кількох кандидатів із різним числом предикторів, використовуючи перехресну перевірку.
9-10	**Тема 5.** Системний підхід до нормалізації та стандартизації даних у задачах машинного навчання. *Робота №5: Методи нормалізації та стандартизації даних в завданнях МН.* **Завдання:** Реалізувати загальний алгоритм до нормалізації та стандартизації даних у задачах машинного навчання. Визначити тип розподілу даних для кожної ознаки та виконати перевірку на нормальність. При необхідності виконати перетворення Бокса-Кокса, або Йео-Джонсона. Виконати перевірку на наявність викидів в наборі даних. Обрати по алгоритму необхідний тип нормалізації і реалізувати його. Підтвердити виконання нормалізації звітом по описової статистиці.
11-12	**Тема 6:** Методи та алгоритми згладжування даних. Системне використання методів фільтрації даних у задачах машинного навчання. *Робота №6: Використання методів та алгоритмів згладжування даних в процедурах аналізу і попередньої обробки даних. Системне використання методів фільтрації даних у задачах машинного навчання.* **Завдання:** 1) Розробити алгоритми згладжування даних (адаптивного кускове згладжування по найближчих сусідах методом найменших квадратів, згладжування з використанням гаусового ядра, медіанного згладжування, експоненціального згладжування та ковзного середнього у вікні). Проаналізувати ефективність використання алгоритмів кожного типу. 2) Побудувати модель процесу, до якого застосовується алгоритм фільтрації. Скористайтесь моделлю процесу АР(р), параметри якої наведені в таблиці згідно варіанту роботи. Порівняти алгоритм оптимальної фільтрації для вільної динамічної системи та алгоритм для системи з детермінованими і випадковими вхідними сигналами.
13-14	**Тема 7:** Інформаційна технологія ймовірнісне-статистичного аналізу та попередньої обробки даних. *Робота №7:Розробка ІТ аналізу та попередньої обробки даних.* **Завдання:** Розробити і реалізувати послідовність методів і підходів аналізу та . попередньої обробки заданого за варіантом набору даних.
15	***Підсумкове тестове опитування.***

4.3. Завдання для самостійної роботи

КАРТА САМОСТІЙНОЇ РОБОТИ

з дисципліни *«Інтелектуальні технології аналізу та попередньої обробки даних»*
кількість годин СРС згідно з навчальним планом ___*75*_____

Види самостійної роботи	Трудомісткість (годин)*	Планові терміни виконання	Форми контролю	Максимальна кількість балів
1	***2***	***3***	***4***	***5***
Денна форма навчання				
10 семестр				
І. О б о в ' я з к о в і				
Види робіт на семінарських (практичних, лабораторних) заняттях				
Опрацювання конспекту лекцій	10	На протязі семестру	Опитування	1
Робота з рекомендованою літературою, наявною в Інтернет-мережі	10	На протязі семестру	Опитування	1
За виконання завдань самостійного опрацювання та інших завдань				
1. Виконання додаткової практичної роботи будь-якого розділу.	25	На протязі семестру	Звіт	3
Разом балів за обов'язкові види СРС				**5**
ІІ. В и б і р к о в і				
За виконання творчих завдань для самостійного опрацювання (один варіант із переліку інд. завдань)				
Аналіз та обробка даних за додатковим варіантом.	30		Звіт	4
Разом балів за вибіркові види СРС				**4**

Вибіркові види самостійної роботи

Фіксований перелік тем для виконання індивідуальних завдань з дисципліни у семестрі студентам не пропонується. Теми обираються студентами самостійно та є поглибленням знань про інтелектуальні технології аналізу та попередньої обробки даних, які розглядаються в межах дисципліни. Крім того, можуть бути розглянутими деякі специфічні використання методів попередньої обробки даних в галузях науки, освіти, промисловості, у медицині або спорті тощо.

Теми індивідуальних завдань узгоджуються з викладачем протягом семестру, до початку залікового тижня. Теми інформаційних повідомлень співпадають з темами та основними питаннями, які розглядаються на лекціях. В інформаційних повідомленнях також можуть розглядатись новітні засобі та методи в сфері моделювання обчислювальних систем.

4.4. Забезпечення освітнього процесу

- Ноутбук, проектор, екран.
- Комплект слайд-презентацій по курсу.
- Програмне забезпечення для демонстрацій слайд-презентацій.
- Мови та програмні середовищ: R Studio, R, Python.

5. Підсумковий контроль

Перелік питань для підготовки до іспиту

1. Що таке попередня обробка даних?
2. Дайте характеристику даним, які використовуються для моделювання, за ступеню структурованості.
3. Дайте характеристику елементам структурованих даних.
4. Дайте характеристику «прямокутним» даним.
5. Дайте характеристику базової таблиці аналітики (склад, вимоги).
6. Звіт про якість даних. Що це?
7. У чому полягають мети дослідження даних?
8. У чому полягає знайомство з даними на базі звіту про якість даних (безперервні ознаки)?
9. У чому полягає знайомство з даними на базі звіту про якість даних (категоріальні ознаки)?
10. Дайте характеристику особливостям розподілів даних. Приклади, висновки для подальшого використання.
11. Дайте характеристику CRISP-DM методології. Склад, призначення.
12. Дайте характеристику основним завданням машинного навчання (регресія, класифікація).
13. Які процедури входять до первинного аналізу даних? Характеристика.
14. Які процедури входять до підготовки даних? Характеристика.
15. Дайте характеристику наявності відсутніх значень в наборах даних. Причини, густина.
16. Опишіть процес обробки відсутніх значень у наборах даних. Видалення ознак і екземплярів ознак.
17. Опишіть процес обробки відсутніх значень у наборах даних. Відновлення даних. Методи відновлення.
18. Опишіть особливості Загальної схеми обробки пропусків в даних.
19. У чому полягає процес визначення та ідентифікація пропусків в даних. Які відомі вам підходи?
20. У чому полягає процес дослідження закономірностей появи відсутніх значень?
21. У чому полягає процес формування набору даних без відсутніх значень?
22. Які підходи реалізуються при прогнозуванні відсутніх значень?
23. Дайте характеристику методу *mice* (метод Гіббса) при заповненні пропусків у вибірці даних.
24. Дайте характеристику методам заповнення пропусків в часових рядах.
25. Дайте характеристику різниці в заповненні пропусків на основі лінійної та стохастичної регресії.
26. Дайте характеристику методу *LOCF (last observation carried forward)* при заповненні пропусків у часових рядах.
27. Дайте характеристику аномальним значенням в наборах даних (шум, викиди).
28. Опишіть особливості Загальної схеми методу ідентифікації та обробки викидів в наборі даних.
29. Опишіть відомі вам методи ідентифікації викидів.
30. Дайте характеристику ESD-тесту (*Generalized Extreme Studentized Deviate*) для виявлення одного або декілька викидів в даних.
31. Дайте характеристику методу Ірвіна для виявлення аномальних значень рівнів часового ряду.

32. Дайте характеристику методам машинного навчання для ідентифікації та обробки викидів в наборі даних.
33. Дайте характеристику методам обробки викидів на основі використання затискного перетворення.
34. Для чого необхідна нормалізація даних? Що таке нормалізація та стандартизація даних?
35. Дайте характеристику поняттю «Розумна нормалізація даних».
36. Алгоритм нормалізації і стандартизації даних на етапах попередньої підготовки даних в завданнях МН.
37. У чому полягає процес аналізу задач МН та методу моделювання на необхідність нормалізації даних?
38. Пояснить вплив розподілу даних на вибір методу нормалізації.
39. Пояснить вплив викидів в даних на вибір методу нормалізації.
40. Пояснить різницю при використання лінійних та нелінійних методів нормалізації.
41. Пояснить особливість методів лінійної нормалізації.
42. Пояснить особливість методів нелінійної нормалізації.
43. Дайте характеристику алгоритму Розумної нормалізації даних.
44. Ознаки для ML. Визначення та види.
45. Як відбувається генерація ознак?
46. Що таке відбір ознак і навіщо він потрібний?
47. Як відбирати ознаки з набору даних для подальшого моделювання? Методи Feature Selection. Методи фільтрації.
48. Як відбирати ознаки з набору даних для подальшого моделювання? Методи *Feature Selection*. Обгорткові методи.
49. Як відбирати ознаки з набору даних для подальшого моделювання? Методи *Feature Selection*. Вбудовані методи.
50. Як відбирати ознаки з набору даних для подальшого моделювання? Прямий відбір.
51. Як відбирати ознаки з набору даних для подальшого моделювання? Зворотний відбір.
52. Як відбирати ознаки з набору даних для подальшого моделювання? Рекурсивне виключення ознак.
53. Для чого використовується регуляризація в процедурі відбору ознак? Типи регуляризації.
54. Пояснить алгоритм «Відбір ознак» до ML-моделі (метод *Backward Elimination* - Зворотне усунення).
55. Алгоритм виявлення та видалення "непотрібних" предикторів.
56. Для чого використовується згладжування даних? Головна мета.
57. Навіщо потрібне згладжування даних? Які завдання виконуються при згладжуванні даних?
58. Як згладжування даних або відсутність згладжування впливає на результати моделювання?
59. Для чого згладжувати дані при розв’язанні задач машинного навчання?
60. Дайте характеристику методам згладжування даних.
61. Дайте характеристику методу ковзного середнього для згладжування даних.
62. Дайте характеристику методу лінійної регресії з адаптивною шириною смуги для згладжування даних.
63. Дайте характеристику методам експоненційного згладжування даних.
64. Особливості реалізації методів експоненційного згладжування за допомогою мови R.
65. Особливості методів фільтрації даних в процесі попередньої обробки.
66. Дайте характеристику методам цифрової фільтрації даних.
67. Дайте характеристику методам оптимальної фільтрації даних.
68. Дайте характеристику методам ймовірнісної фільтрації даних.
69. Обгрунтуйте необхідність візуалізації на етапі аналізу та попередньої підготовки даних для моделювання.
70. Особливості аналізу категоріальних даних.

Типові задачі для розв'язання

Тема: Методи нормалізації та стандартизації даних в завданнях МН.

Завдання: 1. Завантажити початковий набір даних. Проаналізуйте описову статистику, а саме min, max та range для кожної ознаки. Зробіть висновки на необхідність нормалізації.

2. Визначити тип розподілу даних для кожної ознаки та виконати перевірку на нормальність. (Тип розподілу визначати за допомогою графічних методів: квантильних Q-Q графіків та гістограм. Перевірку на нормальність здійснювати за допомогою формальних тестів: тест Шапіро-Уілка, непараметричного критерія і тест Андерсона-Дарлінга, теста Крамера фон Мізеса та тест Лілієфорса.)

3. Якщо розподіл відрізняється від нормального, то виконати перетворення Бокса-Кокса, або Йео-Джонсона. Повторити аналіз описової статистики. Потім повторити п.2. Якщо після цього ви не отримаєте нормальний розподіл, то необхідно використати методи нелінійної нормалізації.

4. При нормальному розподілі ознаки, або близькому до нормального виконайте перевірку на наявність в наборі викидів. Якщо викиди існують, тоді використовуйте методи нелінійної нормалізації. Якщо викидів немає, тоді використовуйте методи лінійної нормалізації.

5. Підтвердити виконання нормалізації звітом по описової статистиці.

До звіту обов'язково додати лістинг коду виконання усіх завдань. Варіанти наборів даних для практичної роботи знаходятся в окремому файлі, якій завантажено в систему Moodle.

«0» варіант залікового білету з зазначенням максимальної кількості балів за кожне виконане завдання

№	Питання білета	Максимальна кількість балів
1	Обґрунтуйте необхідність використання методів фільтрації даних в задачах машинного навчання. Назвіть основні цілі фільтрації даних. Охарактеризуйте основні типи фільтрів.	10
2	Пояснить різницю між лінійними та нелінійними методами нормалізації даних. Наведіть приклади нормалізацій різних типів.	10
3	Обґрунтуйте місці попередньої обробки структурованих даних при розв'язанні задач машинного навчання. Наведіть основні процедури попередньої обробки даних і їх послідовність.	10
	Всього	**30**

6. Критерії оцінювання та засоби діагностики результатів навчання

№	Вид діяльності (завдання)	Максимальна кількість балів
1	Практична робота №1	7
2	Практична робота №2	7
3	Практична робота №3	7
4	Практична робота №4	7
5	Практична робота №5	7
6	Практична робота №6	7
7	Практична робота №7	7
8	Практична робота №8	7
9	Виконання тестового завдання	5
10	Виконання завдань самостійної роботи студента	9
11	Разом за семестр	**70**
12	Залік	30
	Всього	**100**

Критерії оцінювання лабораторних/практичних/індивідуальних/робіт/ доповідей/проєктів

Максимальна кількість балів – студент з високою якістю самостійно виконав весь обсяг робіт, відповідає на всі питання, пов'язані з виконаними роботами, та робить додаткові розрахунки, які йому пропонує викладач. У викладача немає претензій щодо реалізації та вимог до виконання роботи.

70%-99% від максимальної кількості балів – студент з достатньою якістю виконав всі завдання, але в процесі роботи він робив деякі помилки, які, після вказування на них викладачем, самостійно виправляв. На деякі питання він відповідає з похибкою. Запропоновані викладачем додаткові розрахунки робить з деякою потугою. Не всі вимоги до виконання роботи дотримані.

40%-69% від максимальної кількості балів – студент самостійно виконав всі роботи, але якість реалізації недостатня (помилки при розрахунках, не всі вимоги до роботи дотримані). На питання щодо виконання робіт відповідає не зовсім чітко. Є помилки при відповідях.

1%-39% від максимальної кількості балів – студент самостійно виконав не всі роботи, при цьому якість реалізації недостатня (помилки при розрахунках, не дотримується вимог до оформлення роботи). На питання щодо виконання робіт відповідає не чітко. Є грубі помилки при відповідях.

0 балів – студент не виконав весь обсяг робіт, або виконав з грубими помилками. Він має проблеми з розрахунками, не знає теоретичного матеріалу, програмна реалізація не відповідає поставленим вимогам.

При отриманні незадовільної оцінки студент має право виправити всі помилки або виконати нові варіанти завдань, якщо викладач невпевнений, що студент виконав їх самостійно. Такий варіант пропонується, коли студент має багато пропусків занять.

Критерії оцінювання для досягнення максимальної кількості балів при виконанні тестових завдань

Оцінювання тестових завдань студентів по дисципліні «Інтелектуальні технології аналізу та попередньої обробки даних» проводиться по 5-бальній шкалі.

Тести допомагають отримати об'єктивніші оцінки рівня знань, умінь, навиків, перевірити відповідність вимог до підготовки студентів заданим стандартам, виявити пропуски в підготовці студентів.

Виходячи з технологічності процедури тестування відповіді кодуються двійковим кодом: **1** – істинно і **0** – помилково, і у такому вигляді можуть поступати в сучасні автоматизовані системи обробки інформації. Також може використовуватись відсоткове кодування відповіді (згідно з кількістю правильних відповідей у разі можливості надання декількох відповідей) відповідно до вагового коефіцієнту кожної відповіді з усіх можливих.

Тестування з дисципліни «Інтелектуальні технології аналізу та попередньої обробки даних» студенти можуть проходити як одноразово, так і необмежену кількість разів, поки не досягнуть результату, який відповідає їх уяві про власні знання. Чим буде більше наполегливим та зацікавленим студент - тім скоріше він досягне найвищого рівня професійних знань та вмінь. Така наполегливість також є складовою мети дисципліни.

Тестові бали переводяться в традиційну систему оцінок. Наприклад, якщо випробовуваний виконав більше 90 % завдань, то він отримує оцінку "відмінно" (**5** балів), що вирішив від 75 до 90 % завдань "добре" (**4** бали), від 50 до 75 % – "задовільно" (**3** бали). Якщо студент виконав менш ніж 50% завдань, то він обов'язково повинен перескласти тестові завдання для отримання допуску до іспиту.

7. Рекомендовані джерела інформації

7.1. Основні:

1. Bidyuk P., Kalinina I., Gozhyj A. An Approach to Identifying and Filling Data Gaps in Machine Learning Procedures. Lecture Notes on Data Engineering and Communications Technologies (Switzeland). 2022. Vol. 77, pp. 164-176. (ISSN: 2367- 4512)
2. Gozhyj A. P., Kalinina I. A., Bidyuk P. I. Systematic use of nonlinear data filtering methods in forecasting tasks. Applied Aspects of Information Technology 2023; Vol.6 No.4: 345–361 DOI: https://doi.org/10.15276/aait.06.2023.23. UDC 004.852:004.6.
3. Kalinina, I., Bidyuk, P., Gozhyj, A., Gozhyi, V., Nechakhin, V. (2025). Approach to Identification of Anomalous Values in Analysis Tasks and Data Pre-processing. In: Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Data Engineering, Computational Intelligence, and Decision-Making, Vol. 2. ISDMCI 2024. Lecture Notes on Data Engineering and Communications Technologies, vol 244. Springer, Cham. https://doi.org/10.1007/978-3-031-88483-2_6
4. Kalinina, I., Gozhyj, A., Bidyuk, P., Gozhyi, V., Korobchynskyi, M., Nadraga, V. (2025). A Systematic Approach to Data Normalization and Standardization in Machine Learning Problems. In: Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Data Engineering, Computational Intelligence, and Decision-Making, Volume 2. ISDMCI 2024. Lecture Notes on Data Engineering and Communications Technologies, vol 244. Springer, Cham. https://doi.org/10.1007/978-3-031-88483-2_11
5. Kalinina, I., Gozhyj, A., Vysotska V., Malakhov E., Gozhyj V., Tregubova I. System Methodology of Data Analysis and Preprocessing for Solving Classification Problems. Conference: 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT). Lviv, Ukraine (2024) DOI:10.1109/CSIT65290.2024.10982630. URL: https://ieeexplore.ieee.org/document/10982630
6. Гороховатський В.О., Творошенко І.С. Методи інтелектуального аналізу та оброблення даних: навч. посібник. – Харків: ХНУРЕ, 2021. – 92 с. URL: https://cutt.ly/U2DbE5v
7. Іванов С.М., Максишко Н.К., Бречко Д.О. Інтелектуальний аналіз даних: конспект лекцій/ С.М. Іванов, Н.К Максишко Н.К., Д.О. Бречко, – Запоріжжя: ЗНУ, 2020, 156 с.
8. Литвин В.В., Нікольський Ю.В., Пасічник В.В. Аналіз даних та знань. Навчальний посібник. – Магнолія, 2021. – 276 с.
9. Kondruk N.E., Malyar M.M. Analysis of Cluster Structures by Different Similarity Measures. Vol. 57, Issue 3, pp. 436–441, (2021) URL: https://doi.org/10.1007/s10559-021-00368-4
10. Samanta D., Banerjee A. Computationally Intensive Statistics for Intelligent/ D. Samanta, A.Banerjee , – Springer, 2021, 218 p.

7.2. Додаткові:

1. Бідюк П.І., Калініна І.О., Гожий О.П. Байєсівський аналіз даних: [монографія]. – Херсон: Книжкове видавництво ФОП Вишемірський В.С., 2021. – 208 с.
1. Гожий О.П., Калініна І.О., Нечахін В.В. Інтелектуальні технології в керуванні гібридними енергетичними системами: [монографія]. Херсон: Книжкове видавництво ФОП Вишемирський В.С., 2021. 200 с.
2. Засоби підготовки та аналізу даних. Лабораторний практикум [Електронний ресурс]: навч. посіб. для студ. спеціальності 113 «Прикладна математика» / А. Ю. Шелестов, Н. М. Куссуль; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 1086 Кбайт). – Київ: КПІ ім. Ігоря Сікорського, 2021. – 31 с. URL: https://ela.kpi.ua/bitstream/123456789/43491/1/Shelestov_Zasoby-pidhotovky-ta-analizu-danykh_LabPrakt.pdf
3. Попередня обробка та аналіз даних: лабораторний практикум для студ. спеціальності 113 «Прикладна математика» /Уклад.: Н. Е. Кондрук. Ужгород: УжНУ, 2023. – 41 с
4. Слабоспицький. О. С. Задачі класифікації: навч. посіб. / О. С. Слабоспицький. – К. Видавництво «Людмила», 2020. – 43 с URL: http://csc.knu.ua/media/filer_public/fb/84/fb84d8c5-f845-4215-8bc7-33ad5eb0e80e/slabospitsky_os2020.pdf

5. Штовба С.Д. Machine learning: стартовий курс: електронний навчальний посібник / Штовба С.Д., Козачко О.М. – Вінниця: ВНТУ, 2020. – 81 с. URL: https://www.researchgate.net/publication/338924246_Machine_Learning_startovij_kurs
6. Akerkar Ranjendra Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data/ Ranjendra Akerkar, – Springer, 2020, 201 p.
7. Anandan R. A Closer Look at Big Data Analytics/ R. Anandan, – Nova Science Publishers. Inc. , 2021, 366 p.
8. Botros S., Tinley J. High Performance MySQL: Proven Strategies for Operating at Scale 4th Edition/ Silvia Botros, Jeremy Tinley, - O'Reilly Media, 2021, 388 p.
9. John D. Kelleher (ed.) Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies. Cambridge, Mass.: 2020. – URL:
10. https://cutt.ly/X2DPsvsSakarkar Gaurav, Patil Gaurav, Dutta Preteek Machine Learning Algorithms Using Python Programming/ Gaurav Sakarkar, Gaurav Patil, Preteek Dutta, - Nova Science Publishers. Inc., 2021, 218 p.
11. Srinivas M., Sucharitha G., Matta A., Chatterjee P. /M. Srinivas, G. Sucharitha, A. Matta, P. Chatterjee Machine Learning Algorithms and Applications: Theory and Applications, -Wiley-Scrivener Publishing, 2021, 368 p.

7.3. Інформаційні ресурси в Інтернет

1. Національний інститут стандартів і технологій (NIST), Стандарти науки про дані та аналізу великих даних: https://bigdatawg.nist.gov/standards/
2. Група спеціальних інтересів з виявлення знань та інтелектуального аналізу даних (SIGKDD) Асоціації обчислювальної техніки (Association for Computing Machinery, ACM): https://www.kdd.org/
3. Процес міжгалузевого стандарту для інтелектуального аналізу даних (CRISP-DM): https://www.datascience-pm.com/crisp-dm-2/
4. Data Mining Group (DMG): https://www.dmg.org/