МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ЗАТВЕРДЖУЮ
Декан факультету комп'ютерних інформаційних технологій
_____ **Ігор ЯКИМЕНКО**
«30» 08 2024 р.

ЗАТВЕРДЖУЮ
Проректор з науково-педагогічної роботи
_____ **Віктор ОСТРОВЕРХОВ**
«30» 08 2024 р.

РОБОЧА ПРОГРАМА

з дисципліни «Виявлення та обробка аномальних даних»

Ступінь вищої освіти - бакалавр
Галузь знань – 12 Інформаційні технології
Спеціальність – 122 Комп'ютерні науки
Освітньо-професійна програма «Штучний інтелект»

Кафедра прикладної математики

Форма навчання	Курс	Семестр	Лекції, (год.)	Лабора. заняття, (год.)	ІРС, (год.)	Тренінг, (год.)	Самост. робота студ. (год.)	Разом, (год.)	Залік, (сем.)
Денна	3	6	30	30	4	8	78	150	6

30.08.2024

Тернопіль – ЗУНУ
2024

Робочу програму склав доцент кафедри прикладної математики, канд. фіз.-мат. наук Андрій АЛІЛУЙКО

Робоча програма затверджена на засіданні кафедри прикладної математики, протокол № 1 від 26.08.2024 р.

Завідувач кафедри Олеся МАРТИНЮК

Розглянуто та схвалено групою забезпечення спеціальності 122 Комп'ютерні науки, протокол № 1 від 30.08. 2024 р.

Голова групи
забезпечення спеціальності Мирослав КОМАР

Гарант ОПП Василь КОВАЛЬ

СТРУКТУРА РОБОЧОЇ ПРОГРАМИ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
«Виявлення та обробка аномальних даних»

1. Опис дисципліни «Виявлення та обробка аномальних даних»

Дисципліна «Виявлення та обробка аномальних даних»	**Галузь знань, спеціальність, освітньо-професійна програма, ступінь вищої освіти**	**Характеристика навчальної дисципліни**
Кількість кредитів ECTS – 5	Галузь знань – 12 Інформаційні технології	**Статус дисципліни**: вибіркова **Мова навчання:** українська
Кількість залікових модулів – 4	Спеціальність – 122 Комп'ютерні науки	Рік підготовки: 3 Семестр: 6
Кількість змістових модулів – 2	Освітньо-професійна програма – Штучний інтелект	Лекції: 30 год Лабораторні заняття: 30 год.
Загальна кількість годин – 150	Ступінь вищої освіти – бакалавр	Самостійна робота:78 год Тренінг:8 год Індивідуальна робота: 4 год
Тижневих годин – 10 год, з них аудиторних – 4 год. (лекції – 2 год., лабораторні заняття – 2 год.)		Вид підсумкового контролю – залік

2. Мета і завдання вивчення дисципліни «Виявлення та обробка аномальних даних»

2.1. Мета вивчення дисципліни

Метою викладання дисципліни є формування системи теоретичних знань і практичних навичок застосування сучасних методів виявлення та обробки аномалій в наборах даних великих об'ємів.

2.2. Завдання вивчення дисципліни

Завдання вивчення дисципліни «Виявлення та обробка аномальних даних»: ознайомити студентів із сучасними методами пошуку та опрацювання аномалій даних; поглибити та розширити знання у сфері роботи з даними великих об'ємів; здобути практичні навички побудови та реалізації алгоритмів пошуку аномалій за допомогою середовища Python.

2.3. Результати навчання

В результаті вивчення навчальної дисципліни студент повинен:

знати:

- теоретичні основи пошуку аномалій;
- методи, алгоритми пошуку аномалій, обробки даних з використанням технологій машинного навчання;

- сучасні програмні засоби реалізації технологій обробки даних та пошуку аномалій в них.

вміти:

- обґрунтовано підбирати методи та алгоритми для пошуку аномалій при опрацюванні конкретних наборів даних;
- здійснювати оцінювання якості реалізованих алгоритмів та їх оптимізації;
- вирішувати прикладні задачі щодо оптимізації та аналізу даних, прогнозування із врахуванням виявлених аномалій даних;
- використовувати сучасні програмні засоби для реалізації технологій обробки даних та пошуку аномалій в них.

3. Програма навчальної дисципліни «Виявлення та обробка аномальних даних»

Змістовий модуль 1. Методи математичної статистики для виявлення аномалій

Тема 1. Вступ в виявлення аномальних даних.

Поняття аномалій даних. Класифікація аномалій. Класифікація методів виявлення аномалій.

Тема 2. Ймовірнісні та геометричні методи виявлення аномалій.

Статистичні методи виявлення аномалій. Імовірностні моделі виявлення аномалій. Аналіз екстремальних значень. Виявлення аномалії на основі кута. Глибинні техніки виявлення аномалій

Тема 3. Виявлення аномалій в часових рядах.

Поняття аномалій в даних часових рядів. Статистичне управління процесом для виявлення аномалій. Авторегресійні моделі.

Змістовий модуль 2. Машинне навчання в виявленні та обробці даних

Тема 4. Лінійні методи для виявлення аномалій.

Лінійна регресія. Аналіз головних компонент (PCA). Однокласові опорні векторні машини (one-class SVMs).

Тема 5. Виявлення аномалій методами на основі близькості.

Метод k-найближчих сусідів (KNN) при виявленні аномалій. Метод кластеризації k-середніх при виявленні аномалій. Метод локального коефіцієнта викиду (LOF) при виявленні аномалій.

Тема 6. Виявлення аномалій в даних великої розмірності.

Особливості виявлення аномалії в даних великої розмірності. Метод підпростору з пакетуванням ознак для виявлення аномалій у багатовимірних наборах даних. Метод ізольованого лісу для виявлення аномалій у багатовимірних наборах даних (Isolation Forest).

Тема 7. Контрольовані методи виявлення аномалій.

Контрольоване виявлення аномалій. Економічне навчання. Адаптивна повторна вибірка.

Тема 8. Оцінка методів виявлення аномалій.

Метричний аналіз методів виявлення аномалій. Аналіз виявлених аномалій різних типів даних. Метрики AUC ROC, Average Precision, Card Precision top-metric при виявленні аномалій.

Тема 9. Глибинне навчання при виявленні аномалій.

Застосування штучної нейронної мережі при виявленні аномалії. Архітектура автокодувальника. Застосування автокодувальника для аналізу аномалій. Послідовні моделі та репрезентативне навчання.

4. Структура залікових кредитів з дисципліни «Виявлення та обробка аномальних даних»

Структура залікового кредиту

Тема	Кількість годин					
	Лекції	**Лабораторні заняття**	**ІРС**	**Тренінг**	**СРС**	**Контрольні заходи**
Змістовий модуль 1. Методи математичної статистики для виявлення аномалій						
Тема 1. Вступ в виявлення аномальних даних	2	-	2	3	2	Поточне опитування
Тема 2. Ймовірнісні та геометричні методи виявлення аномалій	2	4			8	
Тема 3. Виявлення аномалій в часових рядах	4	4			10	
Змістовий модуль 2. Машинне навчання в виявленні та обробці даних						
Тема 4. Лінійні методи виявлення аномалій	4	4	2	5	10	Поточне опитування
Тема 5. Виявлення аномалій методами на основі близькості	4	2			10	
Тема 6. Виявлення аномалій в даних великої розмірності	4	4			10	
Тема 7. Контрольовані методи виявлення аномалій	2	4			10	
Тема 8. Оцінка методів виявлення аномалій	4	4			8	
Тема 9. Глибинне навчання при виявленні аномалій	4	4			10	
Разом	30	30	**4**	**8**	78	

5. Тематика лабораторних занять

Лабораторна робота 1, 2

Тема. Ймовірнісні та геометричні методи виявлення аномалій

Мета: Виробити навики виявлення аномалій даних статистичними, ймовірнісними та геометричними методам.

Питання для обговорення:

1. Статистичні тести: Z-оцінка, модифікована Z-оцінка, тест Грабса, відстань Махаланобіса.
2. Аналіз екстремальних значень: блокові максимуми (block maxima), пік над порогом (peak over threshold).
3. Виявлення аномалії на основі кута (Angle-Based Anomaly Detection).
4. Глибинна техніка виявлення аномалій (Depth-based techniques).

Лабораторна робота 3, 4

Тема. Виявлення аномалій в часових рядах

Мета: Навчити застосовувати числові характеристики числових рядів для виявлення аномалій. Виробити навички побудови ARIMA моделей для вивчення особливостей часових рядів.

Питання для обговорення:

1. Статистичне управління процесом для виявлення аномалій.
2. ARIMA моделювання для пошуку незвичної поведінки в даних часових рядів.

Лабораторна робота 5, 6

Тема. Лінійні методи виявлення аномалій

Мета: Набути практичний досвід роботи з моделями лінійної регресії, з PCA, з однокласними SVM при виявленні аномалій.

Питання для обговорення:

1. Лінійна регресія. Аналіз головних компонент (PCA).
2. Однокласові опорні векторні машини (one-class SVMs).
3. Порівняння результативності застосування методів.

Лабораторна робота 7

Тема. Виявлення аномалій методами на основі близькості

Мета: Навчитися застосовувати методи KNN, K-Means, LOF для виявлення аномалій.

Питання для обговорення:

1. Метод k-найближчих сусідів (KNN) при виявленні аномалій.
2. Метод кластеризації k-середніх (K-Means) при виявленні аномалій.
3. Метод локального коефіцієнта викиду (LOF) при виявленні аномалій.

Лабораторна робота 8, 9

Тема. Виявлення аномалій в даних великої розмірності

Мета: Розуміти проблеми виявлення аномалій в даних великої розмірності. Володіти метод підпростору з пакетуванням ознак та методом ізольований ліс (Isolation Forest) для виявлення аномальних даних.

Питання для обговорення:

1. Метод підпростору з пакетуванням ознак для виявлення аномалій у багатовимірних наборах даних.
2. Метод ізольованого лісу для виявлення аномалій у багатовимірних наборах даних (Isolation Forest).

Лабораторна робота 10, 11

Тема. Контрольовані методи виявлення аномалій

Мета: Навчитися застосовувати методи контрольованого виявлення аномалій. Отримати практичний досвід проведення економічного навчання, адаптивної повторної вибірки та методів посилення при виявленні аномалій.

Питання для обговорення:

1. Контрольоване виявлення аномалій.

2. Економічне навчання (Cost-sensitive learning).
3. Адаптивна повторна вибірка (adaptive resampling).
4. Методи підсилення (boosting methods).

Лабораторна робота 12, 13

Тема. Оцінка методів виявлення аномалій

Мета: Навчити оцінювати методи виявлення аномалій.

Питання для обговорення:

1. Оцінка виявлених аномалій.
2. Показники точності виявлення аномалій.
3. Застосування метрик AUC ROC, Average Precision, Card Precision top-metric при виявленні шахрайств .

Лабораторна робота 14, 15

Тема. Глибинне навчання при виявленні аномалій

Мета: Навчити застосовувати методи глибинного навчання при виявленні аномалій.

Питання для обговорення:

1. Застосування штучної нейронної мережі FNN при виявленні карткових шахрайств.
2. Використання автоенкодеру при виявленні аномалій.
3. Послідовні моделі та репрезентативне навчання при виявленні аномалій.

6. Тренінг з дисципліни

Мета тренінгу з дисципліни «Виявлення та обробка аномальних даних» – оволодіння навичками застосування методів обробки та пошуку аномальних даних різних галузей.

Проведення тренінгу дозволяє: забезпечити практичне засвоєння теоретичних знань, отриманих у процесі вивчення дисципліни «Виявлення та обробка аномальних даних»;

Завдання 1.

- завантажити набір даних із сайтів: https://www.kaggle.com/; https://www.timeseriesclassification.com/dataset.php, або інших сайтів. Здійснити його аналіз та за необхідності їх попереднє опрацювати.
- використати хоча б два методи математичної статистики для відшукання аномалій відповідно до структури підготовленого набору даних.
- для отриманих результатів здійснити порівняння методів, оцінити їх ефективність. Зробити необхідні висновки.

Завдання 2.

- завантажити набір даних із сайтів: https://www.kaggle.com/; https://www.timeseriesclassification.com/dataset.php, або інших сайтів. Здійснити його аналіз та за необхідності їх попереднє опрацювати.
- використати хоча б два методи машинного навчання для відшукання аномалій відповідно до структури підготовленого набору даних.

– для отриманих результатів здійснити порівняння методів, оцінити їх ефективність. Зробити необхідні висновки.

Загальна оцінка студента за роботу під час тренінгу визначається як середнє арифметичне з оцінок, отриманих за виконання завдань на тренінгу.

7. Самостійна робота студентів

З метою засвоєння дисципліни «Виявлення та обробка аномальних даних» студенти повинні володіти значним обсягом інформації, частину якої вони отримують і опрацьовують шляхом самостійної роботи. Виконання самостійної роботи полягає в ознайомлені студентами із Python бібліотекою PyOD [9]. Студент для виконання самостійно генерує набори даних (одновимірні, багатовимірні, часовий ряд) або можна завантажити із сайтів: https://www.kaggle.com/; https://www.timeseriesclassification.com/dataset.php, або інших сайтів.

Самостійна робота виконується протягом семестру і складається з 13 завдань. Кожне завдання оцінюється від 1 до 100 балів залежно від повноти виконання, кількості допущених помилок.

№ п/п	**Алгоритми виявлення аномалій** PyOD
1	Angle-Based Outlier Detection (ABOAD, FastABOD)
2	Minimum Covariance Determinant (використовуйте відстані Махаланобіса як оцінки аномалій) (MCD)
3	Principal Component Analysis (PCA)
4	One-Class Support Vector Machines (OCSVM)
5	Local Outlier Facto (LOF)
6	k Nearest Neighbors (використовуйте до k-го найближчого сусіда як аномалію) (kNN)
7	Average kNN (використовуйте середню відстань до k найближчих сусідів як аномалію)
8	Median kNN (використовуйте середню відстань до k найближчих сусідів як викид)
9	Subspace Outlier Detection (SOD)
10	Isolation Forest (IForest)
11	Extreme Boosting Based Outlier Detection (XGBOD)
12	Fully connected AutoEncoder (використовуйте помилку реконструкції як показник аномалії)
13	Feature Bagging (пакетування ознак)

8. Методи навчання

У навчальному процесі застосовуються: лекції, лабораторні заняття, консультації, самостійна робота, метод опитування, тестування.

9. Засоби оцінювання та методи демонстрування результатів навчання

У процесі вивчення дисципліни «Виявлення та обробка аномальних даних» використовуються наступні засоби оцінювання та методи демонстрування результатів навчання:

- поточне опитування та тестування;
- оцінювання результатів модульних робіт;
- оцінювання виконання завдань на тренінгу;
- оцінювання результатів самостійної роботи.

10. Критерії, форми поточного та підсумкового контролю

Підсумковий бал (за 100-бальною шкалою) з дисципліни «Виявлення та обробка аномальних даних» визначається як середньозважена величина, залежно від питомої ваги кожної складової залікового кредиту:

Модуль 1		**Модуль 2**		**Модуль 3**	**Модуль 4**
20%	20%	20%	20%	5%	15%
Поточне оцінювання	Модульний контроль 1	Поточне оцінювання	Модульний контроль 2	Тренінги	Самостійна робота
Визначається як середнє арифметичне з оцінок, отриманих за виконання та захист лабораторних робіт. Опитування проводиться з тем 1-5	Тестові завдання (10 тестів по 2 бали за тест) – макс. 20 балів Задача 1 – макс. 40 балів Задача 2 – макс. 40 балів	Визначається як середнє арифметичне з оцінок, отриманих за виконання та захист лабораторних робіт. Опитування проводиться з тем 6-9	Тестові завдання (10 тестів по 2 бали за тест) – макс. 20 балів Задача 1 – макс. 40 балів Задача 2 – макс. 40 балів	Визначається як середнє арифметичне з оцінок, отриманих за виконання завдань на тренінгу	Визначається як середнє арифметичне з оцінок, отриманих за виконання завдань самостійної роботи

Шкала оцінювання:

За шкалою ЗУНУ	**За національною шкалою**	**За шкалою ECTS**
90–100	Відмінно	A (відмінно)
85-89	Добре	B (дуже добре)
75–84		C (добре)
65–74	задовільно	D (задовільно)
60-64		E (достатньо)
35–59	незадовільно	FX (незадовільно з можливістю повторного складання)
1–34		F (незадовільно з обов’язковим повторним курсом)

11. Інструменти, обладнання та програмне забезпечення, використання яких передбачає навчальна дисципліна

№	Найменування	Номер теми
1.	Комунікаційне програмне забезпечення (Zoom) для проведення занять у режимі онлайн (за необхідності)	1–9
2.	Комунікаційна навчальна платформа (Moodle) для організації дистанційного навчання (за необхідності)	1–9
3.	Python	1–9
4.	Мультимедійне обладнання	1–9

РЕКОМЕНДОВАНІ ДЖЕРЕЛА ІНФОРМАЦІЇ

Основна література

1. Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)* 54, 2021. 3. P. 1–33.
2. Julien Lesouple, Cédric Baudoin, Marc Spigai, Jean-Yves Tourneret. Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*. 2021. Vol. 149. P. 109-119.
3. Mohammad Braei and Sebastian Wagner. Anomaly detection in univariate time-series: A survey on the state-of-the-art. arXiv preprint arXiv:2004.00433, 2020.
4. Документація по пакету PYOD – URL: https://pyod.readthedocs.io/en/latest
5. Гавриленко С.Ю., Зозуля В.Д. Дослідження методів виявлення аномалій на етапі попередньої обробки даних. *Системи управління, навігації та зв'язку*. 2022. Вип. 1 (67). с. 52-56.
6. Завгородній В.В. Завгородня Г.А. Валявська Н.О. Герасименко О.О. Калюжний О.В. Степовий А.В. Пошук аномалій у даних за допомогою машинного навчання. *Вчені записки ТНУ імені В.І. Вернадського*. Серія: Технічні науки. 2022. Т. 33 (72). № 3 с. 39-43.

Додаткова література

7. Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal*. 2021. 30, 6, P. 909–931.
8. Anomaly Detection. Intel course – URL: https://www.intel.com/content/www/us/en/developer/learn/course-anomaly-detection.html