Силабус курсу

Інтелектуальний аналіз даних

Ступінь вищої освіти – бакалавр
Освітньо-професійна програма «Професійна освіта (Цифрові технології)»

Рік навчання: 3, **Семестр:** 5

Кількість кредитів: 5 **Мова викладання:** українська

Керівник курсу

ПІП к.е.н., доцент **Андрушків Роман Юрійович**

Контактна інформація andrushkiv81roman@gmail.com, +380677910761

Опис дисципліни

Процес інтелектуального аналізу даних полягає у виділенні потрібної інформації з набору даних та її структуризація у доступну форму для подальшого використання. Інтелектуальний аналіз даних - це один із визначальних кроків у процесі виявлення релевантної інформації. Наступні кроки включають оцінку закономірностей, подібну до процесу аналізу (на цьому етапі відбувається інтерпретація виявлених закономірностей і взаємозв'язків), та інтеграцію інформації, подібну до звіту про результати, проте вони повинні характеризуватись високою надійністю, а ніж проста інтеграція інформації, щоб ретельно відповідати встановленим вимогам. Як і аналіз даних (Data Mining), пошук необхідної інформації у базах даних (Knowledge Discovery in Data, далі – KDD) є ітеративним процесом. Якщо закономірності, оцінені на етапі інтелектуального аналізу даних, не перетворились у високопродуктивні, даний процес доцільно починати знову з попереднього етапу.

Іншими словами, основна проблема, яку вирішує процес KDD, полягає в перетворенні низькорівневих даних на більш корисні та стислі звіти, абстрактні описи або прогнозні моделі. Саме тому, у даному курсі розглядається як історичне підґрунтя виявлення релевантної інформації у базах даних та інтелектуального аналізу даних, з особливим акцентом на його перетині з іншими суміжними галузями, так і лаконічно розглядаються нещодавні реальні застосування KDD. Так подається визначення KDD та інтелектуального аналізу даних і описується загальний багатоетапний процес KDD, що передбачає використання алгоритмів інтелектуального аналізу даних.

Структура курсу

Години (лек. / сем.)	Тема	Результати навчання	Завдання
4 / 4	Тема 1. Концепції інтелектуального аналізу даних та DataMining	Визначення необхідності аналізу великих, складних, багатих інформацією наборів даних, цілями та основними завданнями процесу інтелектуального аналізу даних. Знати технології інтелектуального аналізу даних та Data Mining, ітераційний характер процесу інтелектуального аналізу даних та його основних кроків, вплив якості даних на процес інтелектуального аналізу данихта концепцію великих даних та науки про дані. Вміти застосовувати алгоритми Data Mining для вирішення реальних завдань аналізу даних, визначення корисних шаблонів та тенденцій у даних.	Питання, практичне завдання

2 / 2	Тема 2. Поняття даних. Типи та формати зберігання даних. Бази даних. СУБД	Засвоєння основних концептуальних поняття з курсу «Інтелектуальний аналіз даних»; засвоєння відмінності Data Mining від класичних статистичних методів аналізу й OLAP-систем, вивчення типів закономірностей, що виявляють Data Mining та класи систем інтелектуального аналізу даних. Вивчення даних, набори даних та їх атрибутів, формати зберігання даних, якісного аналізу даних із використанням Data Mining (DM), систем управління базами даних. Опрацювання п'яти типів шкал вимірювань: номінальної, порядкової, інтервальної, відносної і дихотомічної; висування гіпотез, а також процесом збору, систематизації даних та вимог до СУБД.	Питання, презентація, практичне завдання
2 / 2	Тема 3. Метадані. Класифікація метаданих	Ознайомити з поняттям «метадані» та особливостей роботи з ними. Визначення відмінностей між даними і метаданими. Класифікація метаданих. Формат метаданих. Стандарти W3C, ISO, ANSI тощо.	Питання, практичне завдання
4 / 4	ТЕМА 4. Етапи ІАД. Класифікація методів ІАД	Ознайомити з методами і алгоритмами Data Mining: штучні нейронні мережі, дерева рішень, символьні правила, методи найближчого сусіда і k-найближчого сусіда, метод опорних векторів, байєсовські мережі, лінійна регресія, кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, в тому числі алгоритми k-середніх і k-медіани; методи пошуку асоціативних правил, у тому числі алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних і безліч інших методів. Вміти інтерпретувати та класифікувати технологічні методи Data Mining, як кластерний аналіз, метод найближчого сусіда, метод k-найближчого сусіда, міркування за аналогією.	Питання, презентація, практичне завдання
2 / 2	Тема 5. Задачі Data Mining та їх класифікація. Інформація та знання	Ознайомити з задачами Data Mining, їх класифікація. Методи розв'язання: найближчого сусіда (Nearest Neighbor), k-найближчого сусіда (k-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks). засвоїти властивості методів інтелектуального аналізу даних: кластеризація (Clustering); асоціація (Associations); послідовність (Sequence); прогнозування (Forecasting); визначення відхилень або викидів (Deviation Detection). Вміти класифікувати задачі інтелектуального аналізу даних, рівні аналізу, інформації та її властивостей.	Питання, практичне завдання

2 / 2	Тема 6. Задачі Data Mining. Класифікація та кластеризація	Ознайомити із задачами та видами класифікації: допоміжна (штучна) класифікація, природна класифікація, проста та складна класифікація. Опрацювати контрольоване або кероване навчання, маши́нне навчання, навчання з вчителем, навчання без вчителя, навчання із закріпленням. Вміти інтерпретувати одновимірну та багатовимірну класифікацію, процес класифікації та класифікатор, навчальну множину. Конструювати моделі та використання їх. Методи, що застосовуються для розв'язання задач класифікації. Кластеризація та її задачі. Застосування кластерного аналізу. Вивчити задачі класифікації та кластеризації; засвоїти принцип штучної та природної класифікації.	Питання, практичне завдання
4 / 4	Тема 7. Задачі Data Mining. Прогнозування та візуалізація	Ознайомити із задачами прогнозування, пняттям прогнозування і часовими рядами; трендом, сезонністю і циклом; періодом прогнозування; горизонтом прогнозування; інтервалом прогнозування. Встановлення точності прогнозу, видів помилок та прогнозів; візуалізація інструментів Data Mining. Опрацювання методів візуалізації, принципів компонування візуальних засобів, лаконічності, узагальнення й уніфікації, принципів акценту на основних значимих елементах, принципів автономності, структурності, стадійності. Визначення основні тенденції в області візуалізації. Вивчити задачі інтелектуального аналізу даних; засвоїти рівні аналізу Data Mining; засвоїти поняття інформації та вивчити її властивості. Вивчити задачі класифікації та кластеризації; засвоїти принцип штучної та природної класифікації.	Питання, презентація, практичне завдання
4 / 4	Тема 8. Основи аналізу даних	Ознайомити із основами підготовчих етапів процесу Data Mining - традиційний процес та його етапи, а саме: аналіз предметної області; постановка задачі; підготовка даних; побудова моделей; перевірка й оцінка моделей; вибір моделі; застосування моделі; корекція й відновлення моделі. Засвоїти дублювання даних та дублікати, шуми й викиди, очищення даних (data cleaning, data cleansing або scrubbing) та їх етапи: аналіз даних; визначення порядку й правил перетворення даних; підтвердження; перетворення; протитечія очищених даних; інструменти ETL. Вивчити основи інтелектуального аналізу даних; засвоїти, в чому полягає процес очищення даних.	Питання, презентація, практичне завдання

4 / 4	Тема 9. Методи дерев рішень, класифікації та прогнозування	Ознайомити студентів з методом дерева рішень (decision trees), його перевагами, процесом його конструювання, алгоритмом конструювання, а також з його етапами: «побудовою» або «створенням» дерева (tree building) і «скороченням» дерева (tree pruning). Визначення алгоритмів реалізації дерева рішень: CART, C4.5, CHAID, CN2, Newid, Itrule і інші. Вміти інтерпретувати та застосовувати метод опорних векторів (Support Vector Machine – SVM), лінійний SVM, метод «найближчого сусіда». Байєсовська класифікація. Теорема Байєса. Байєсовський класифікатор. Вивчити методи прогнозування та класифікації; засвоїти поняття дерева рішень; вивчити метод опорних векторів; засвоїти метод найближчого сусіда; засвоїти поняття байєсовської класифікації.	Питання, презентація, практичне завдання
2 / 2	Тема 10. Методи кластерного аналізу. Ієрархічні методи	Ознайомити з кластерним аналізом та його задачами, методами: ієрархічними методами кластерного аналізу, ієрархічними агломеративними методами (Agglomerative Nesting, AGNES), ієрархічними дивизимними (ділені) методами (Divisive Analysis, DIANA). Візуалізація дендрограми (dendrogram), міри подібності, квадрату евклідової відстані, Манхеттенска відстані, відстань Чебишева. Характеристика відсотку незгоди, методів об'єднання або зв'язків, методів найближчого сусіда або одиночного зв'язку, методу найбільш віддалених сусідів або повного зв'язку, методу Варда (Ward's method), методу незваженого попарного середнього, методу зваженого попарного середнього, незваженого центроїдного методу, зваженого центроїдного методу (метод зваженого попарного центроїдного усереднення – weighted pair-group method using the centroid average, WPGMC). Вивчити методи кластерного аналізу; засвоїти особливості ієрархічних методів.	Питання, практичне завдання

Літературні джерела

1. Гороховатський В. О. Методи інтелектуального аналізу та оброблення даних: навч. посіб. В. О. Гороховатський, І. С. Творошенко; М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. Харків: ХНУРЕ, 2021. 92 с.
2. Бахрушин В. Є. Методи аналізу даних: Навч. посібник В. Є. Бахрушин. Запоріжжя: КПУ, 2019. 268 с.
3. Інтелектуальний аналіз даних: Комп'ютерний практикум [Електронний ресурс]: навч. посіб. для студ. спеціальності 122 «Комп'ютерні науки та інформаційні технології», спеціалізацій «Інформаційні системи та технології проектування», «Системне проектування сервісів». О. О. Сергеєв-Горчинський, Г. В. Іщенко; КПІ ім. Ігоря Сікорського. Київ: КПІ ім. Ігоря Сікорського, 2020. 73 с.
4. Акіменко В.В. Прикладні задачі інтелектуального аналізу даних (DATA MINING). К.: КНУ ім. Тараса Шевченко, 2020. 152 с.
5. Jiajun, Z., Zong, C., & Xia, R. (2022). Text Data Mining. Springer.
6. Li, B., Yue, L., Jiang, J., Chen, W., Li, X., Long, G., Fang, F., & Yu, H. (Ред.). (2022). Advanced Data Mining and Applications. Springer International Publishing. https://doi.org/10.1007/978-3-030-95408-6
7. Park, L. A. F., Gomes, H. M., Doborjeh, M., Boo, Y. L., Koh, Y. S., Zhao, Y., Williams, G., & Simoff, S. (Ред.). (2022). Data Mining. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8746-5
8. Tan, Y., & Shi, Y. (Ред.). (2022). Data Mining and Big Data. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8991-9
9. Ufuk Baytar, C. (Ред.). (2022). The Future of Data Mining. Nova Science Publishers. https://doi.org/10.52305/kcin5931
10. Cao, J. (2023). E-Commerce Big Data Mining and Analytics. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3588-8
11. Konys, A., & Nowak-Brzezińska, A. (Ред.). (2023). Knowledge Engineering and Data Mining. MDPI. https://doi.org/10.3390/books978-3-0365-6789-1
12. Mai, T. T., Crane, M., & Bezbradica, M. (Ред.). (2023). Educational Data Mining und Learning Analytics. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978- 3-658-39607-7
13. Olson, D. L., & Lauhoff, G. (2023). Deskriptives Data-Mining. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21274-1
14. Shah, K., Shah, N., Sawant, V., & Parolia, N. (2023). Practical Data Mining Techniques and Applications. Auerbach Publications. https://doi.org/10.1201/9781003390220
15. Zhang, H. (2023a). Handbook of Mobility Data Mining, Volume 1: Data Preprocessing and Visualization. Elsevier.
16. Zhang, H. (2023b). Handbook of Mobility Data Mining, Volume 2: Mobility Analytics and Prediction. Elsevier.
17. Zhang, H. (2023c). Handbook of Mobility Data Mining, Volume 3: Mobility Data-Driven Applications. Elsevier.

Політика оцінювання

У процесі вивчення дисципліни «Інтелектуальний аналіз даних» використовуються такі засоби оцінювання та методи демонстрування результатів навчання: поточне опитування, тестування; презентації результатів виконаних завдань; оцінювання результатів модульної контрольної роботи; оцінювання тренінгового завдання; оцінювання результатів самостійної роботи студентів; інші види індивідуальних і групових завдань; екзамен.

Політика щодо дедлайнів і перескладання. Для виконання індивідуальних завдань і проведення контрольних заходів встановлюються конкретні терміни. Перескладання модулів відбувається з дозволу дирекції факультету за наявності поважних причин (наприклад, лікарняний).

Політика щодо академічної доброчесності. Використання друкованих і електронних джерел інформації під час контрольних заходів та екзаменів заборонено.

Політика щодо відвідування. Відвідування занять є обов'язковим компонентом оцінювання. За об'єктивних причин (наприклад, карантин, воєнний стан, хвороба, закордонне стажування) навчання може відбуватись в он-лайн формі за погодженням із керівником курсу з дозволу дирекції факультету.

Оцінювання

Підсумковий бал (за 100-бальною шкалою) з дисципліни «Інтелектуальний аналіз даних» визначається як середньозважена величина в залежності від питомої ваги кожної складової залікового кредиту:

Модуль 1		Модуль 2		Модуль 3	Модуль 4	Модуль 5
10 %	10 %	10 %	10 %	5 %	15 %	40 %
Поточне опитування	Модульний контроль	Поточне опитування	Модульний контроль	Тренінг	Самостійна робота	Екзамен
Оцінюється як середнє арифметичне з оцінок, отриманих по темах 1-5	Рубіжна контрольна робота по темах 1-5. 1.Теоретичні питання (2 питання – макс. по 25 балів). 2. Тестові завдання (5 тестів по 5 балів за тест) – макс. 25 балів 3. Задача 1 – макс. 25 балів	Оцінюється як середнє арифметичне з оцінок, отриманих по темах 6 -10	Підсумкова контрольна робота по темах 6-10. 1.Теоретичні питання (2 питання – макс. по 25 балів). 2. Тестові завдання (5 тестів по 5 балів за тест) – макс. 25 балів 3. Задача 1 – макс. 25 балів	Оцінюється практичне завдання – макс. 100 балів	Сукупність питомої ваги кожної складової: 1. Підготовка презентації – 80%. 2. Захист презентації – 20%.	1. Тестові завдання (10 тестів по 4 бали за тест) – макс. 40 балів. 2. Задача – макс. 40 балів. 3. Теоретичне питання – макс. 20 балів.

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)