СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
«МЕТОДИ ТА ТЕХНІКИ АНАЛІЗУ ВЕЛИКИХ ДАНИХ»

Ступінь освіти	магістр
Галузь знань	Інформаційні технології
Тривалість викладання	3, 4 чверті
Заняття:	ІІ семестр 2020/2021 н.р.
лекції:	1 година
лабораторні заняття:	2 години
Мова викладання	українська

Сторінка курсу в СДО НТУ «ДП»:

Інші додаткові ресурси: https://www.netacad.com/courses/cybersecurity/ccna-security

Кафедра, що викладає Інформаційних технологій та комп'ютерної інженерії

Інформація про викладача:

Викладач:
Кожевніков Антон Вячеславович
ас. кафедри

Персональна сторінка
https://it.nmu.org.ua/ua/HR_staff/prepods/kozhevnykov.php

E-mail:
kozhevnykov.a.v@nmu.one

1. Анотація до курсу

Дані всюди. Важко уявити, скільки нових даних генерується кожен день. Дані можуть бути словами в книзі, вмістом електронної таблиці, зображеннями, відео, аудіо або потоком вимірювань, відправленим з пристрою моніторингу. Необроблені дані мають мало користі. Ми повинні обробити ці дані, а потім інтерпретувати вивід, щоб зробити їх корисним. Ці корисні дані тепер перетворилися в інформацію.

Коли даних так багато, що традиційні способи їх обробки, зберігання та аналізу неможливо використовувати, це називається великими даними. Великі дані вимагають нових методів та інструментів, щоб зробити їх значущими. Цей курс знайомить студентів з цими методами та інструментами, які допоможуть використовувати всю міць великих даних.

2. Мета та завдання курсу

Мета дисципліни – Мета дисципліни – формування умінь та компетенцій щодо методів та інформаційні технології обробки даних, розмір яких перевищує можливості звичайних програмних аналітичних платформ та баз даних по аналізу, зберіганню і управлінню великими даними (Big Data). Реалізація мети вимагає визначення дисциплінарних результатів навчання та адекватний відбір змісту навчальної дисципліни за цим критерієм.

Предметом вивчення дисципліни є технології, методи та засоби обробки і візуалізації великих даних.

Завдання вивчення дисципліни:

- оволодіння основними поняттями інтелектуального аналізу даних;
- ознайомлення з новітніми інформаційними технологіями аналітичної обробки інформації;
- набуття практичних навичок використання методів і засобів інтелектуального аналізу даних.

3. Результати навчання

Студенти

Знають: основні поняття, методи, засоби, моделі та алгоритми інтелектуального аналізу великих даних.

Розуміють: принципи застосування технологій інтелектуального аналізу даних, перспективи і проблеми аналітики великих даних в зв'язку з Інтернетом речей.

Уміють: вільно орієнтуватися на сучасному ринку аналітичних програмних продуктів; використовувати Jupyter Notebooks для створення конвеєра даних для збору, аналізу та візуалізації даних; застосовувати моделі машинного навчання для автоматизації завдань.

Компетенції:

- студент спроможний розрізняти структуровані дані та неструктуровані дані;
- студент спроможний аналізувати дані за допомогою Python і SQLite;

- студент спроможний аналізувати дані, використовуючи основні статистичні методи і методи підготовки даних в Python за допомогою pandas;
- студент спроможний аналізувати дані за допомогою моделей машинного навчання.

4. Структура курсу

1. Вступ
1.1 Мета і завдання дисципліни “ Методи та техніки аналізу великих даних”.
1.2 Сфери застосування Big Data
2. Подання даних та їх попередня обробка
2.1 Визначення.
2.2 Шкали виміру ознак.
2.3 Життєвий цикл даних.
2.4 Попередня обробка даних.
2.5 Алгоритм ZET заповнения пробілів у таблицях даних.
2.6 Метадані. Життєвий цикл метаданих.
3. Прогнозування стохастичних залежностей.
3.1 Регресійний аналіз. Завдання регресійного аналізу.
3.2 Оцінка параметров рівнянь парної регресії.
3.3 Система нормальних рівнянь.
3.4 Основні моделі парної регресії. Критерії оцінки якості моделей парної регресії.
3.5 Багатовимірний лінійний регресійний аналіз. Відбір факторних змінних. Мультиколлінеарність.
3.6 Алгоритм Фаррара-Глобера. Оцінка параметров рівняння множинної лінійної регресії. Стандартизоване рівняння множинної лінійної регресії. Покроковий відбір факторів множинної лінійної регресії
3.7 Аналіз та прогнозування часових рядів. Автокорреляція часових рядів. Стаціонарність часових рядів. AR та MA процеси
4. Методи класифікації даних. Кластерний аналіз
4.1 Огляд методів класифікації даних.
4.2 Завдання кластерного аналізу. Міри відстані та збіжності. Основні метрики.
4.3 Основні методи кластерного аналізу: ієрархічні агломеративні та дивізимні методи, метод K-середніх Мак-Куїна.
4.4 Нечітка кластеризація, метод C-середніх
5. Методи класифікації даних. Дерева рішень
5.1 Структура дерева рішень і цільова функція.
5.2 Критерії обрання атрибута розбиття дерева.
5.3 Критерії припинення розбиття. Відсікання гілок.
5.4 Алгоритм ID3
6. Методи класифікації даних. Нейронні мережі

6.1 Елементи нейронної мережі.
6.2 Функції активації нейрона.
6.3 Основні архітектури мереж.
6.4 Помилка мережі. Навчання мереж.
6.5 Правила Хебба. Правило Відроу-Хоффа.
6.6 Реалізація бінарного класифікатора та предиктора лінійного часового тренду на основі нейронних мереж.

7. Факторний аналіз

7.1 Завдання факторного аналізу.
7.2 Факторне відображення.
7.3 Визначення факторних навантажень.
7.4 Метод головних компонент.

8. Програмні інструменти Big Data.

8.1 Мова Python
8.2 Статистичні функції мови Python.
8.3 Аналіз даних засобами мови: регресійний, кластерний, факторний.
8.4 Реалізація дерев рішень та нейронних мереж засобами мови.

9. Програмні інструменти Big Data.

9.1 Платформа Hadoop
9.2 Екосистема Hadoop.
9.3 Складові платформи: Hadoop Common, YARN, MapReduce, файлова система HDFS. Пісочниця Apache Hadoop

ЛАБОРАТОРНІ ЗАНЯТТЯ

Лабораторна робота 1	Ознайомлення з інтерфейсом користувача аналітичної платформи Deductor Academic. Експорт, імпорт і візуалізація даних
Лабораторна робота 2	Deductor Academic. Сховище даних.
Лабораторна робота 3	Індивідуальна робота за вибором викладача: Очищення даних OLAP-технологія аналізу даних Автокореляційний аналіз часового ряду ABCD-аналіз даних XYZ-аналіз даних Прогнозування часового ряду за допомогою нейронної мережі Регресійне прогнозування часового ряду Класифікація даних за допомогою дерев рішень Кластеризація даних ієрархічними методами Кластеризація даних за допомогою самоорганізуючихся карт Кохонена

	Пошук асоціативних правил
Лабораторна робота 4	MathCAD. Регресійний аналіз
Лабораторна робота 5	MathCAD. Кластерний аналіз.

5. Технічне обладнання та/або програмне забезпечення

Використовуються лабораторна та інструментальна бази випускової кафедри інформаційних технологій та комп'ютерної інженерії, а також комп'ютерне та мультимедійне обладнання:

1. Персональний комп'ютер або ноутбук зі сталим доступом до мережі Інтернет
2. Активований акаунт університетської пошти (student.i.p.@nmu.one) на Офіс365.
3. Дистанційна платформа Moodle. Активний обліковий запис у системі дистанційної освіти Moodle.
4. Програмне забезпечення:
   - платформа Windows 10;
   - Microsoft Office або LibreOffice;
   - інтернет-браузер;
   - аналітична платформа Deductor Academic.

6. Система оцінювання та вимоги

6.1. Навчальні досягнення здобувачів вищої освіти за результатами вивчення курсу оцінюватимуться за шкалою, що наведена нижче:

Рейтингова	Інституційна
90…100	відмінно / Excellent
74…89	добре / Good
60…73	задовільно / Satisfactory
0…59	незадовільно / Fail

6.2. Здобувач вищої освіти може отримати **підсумкову оцінку** з навчальної дисципліни на підставі поточного оцінювання знань за умови, якщо набрана кількість балів з поточного тестування та самостійної роботи складатиме не менше 60 балів.

Поточна успішність складається з оцінок за лекційну частину курсу та лабораторний практикум. Отримані бали додаються і є підсумковою оцінкою за вивчення навчальної дисципліни. Максимально за поточною успішністю здобувач вищої освіти може набрати 100 балів.

Максимальне оцінювання:

Теоретична частина	Лабораторна частина		Разом
	При своєчасному складанні	При несвоєчасному складанні	
50	50	40	100

В рамках курсу передбачено виконання 7 лабораторних робіт. Під час захисту лабораторної роботи здобувач відповідає на запитання стосовно ходу роботи, пояснює послідовність дій, демонструє результати роботи.

За результатами виконання лабораторної роботи здобувачі складають звіт встановленого зразка, який завантажується до системи Moodle у відповідну категорію.

Звіт обов'язково має містити такі структурні компоненти:

- титульний лист;
- номер варіанту, текст завдання;
- скріншоти етапів виконання завдання, посилання на відповідні ресурси, коди програм тощо;
- звіт має бути завантажено у систему впродовж 3 днів після захисту роботи на занятті.

Важливо!!! Всі умови до лабораторних робіт з детальними поясненнями до них представлено на сторінці Moodle. Всі бали за лабораторні роботи фіксуються у журналі оцінок Moodle.

Індивідуальне завдання. У здобувачів вищої освіти є можливість отримати індивідуальне завдання, що дозволить повищить результуючу оцінку з навчальної дисципліни.

Цей вид роботи складається з 2 завдань:

1. Підготувати набір зображень, виданих викладачем відповідно до обраної предметної області здобувача.
2. Сконфігурувати нейронну мережу для класифікації зображень засобами Python.

6.3. Критерії оцінювання теоретичної частини курс.

Під час проведення контрольних заходів наприкінці третьої та четвертої чверті здобувачі вищої освіти складають відповідні тести, кожен з яких складається з 25 питань. На кожне питання надається 4 варіанти відповіді, серед яких лише 1 – вірний. Максимальна оцінка за тест складає 25 балів, максимальна оцінка за теоретичну частину курсу (сума оцінок за 2 тести) – 50 балів. Опитування за тестом проводиться з використанням системи дистанційної освіти Moodle.

6.4. Критерії оцінювання лабораторної роботи.

З кожної лабораторної роботи здобувач вищої освіти отримує 5 запитань з переліку контрольних запитань. Відповідь на питання оцінюється максимально у 2 бал, причому:

– **2 бали** – відповідь правильна:

– **1 бал** – відповідь вірна, але не повна;
– **0, 5 бали -** ; відповідь вірна, але містить неточності та/або помилки;
– **0 балів** – відповідь неправильна.

Максимальна оцінка за лабораторну роботу складає 10 балів. Максимальна оцінка за лабораторний практикум – 50 балів.

7. Політика курсу

7.1. Політика щодо академічної доброчесності

Академічна доброчесність здобувачів вищої освіти є важливою умовою для опанування результатами навчання за дисципліною і отримання задовільної оцінки з поточного та підсумкового контролів. Академічна доброчесність базується на засудженні практик списування (виконання письмових робіт із залученням зовнішніх джерел інформації, крім дозволених для використання), плагіату (відтворення опублікованих текстів інших авторів без зазначення авторства), фабрикації (вигадування даних чи фактів, що використовуються в освітньому процесі). Політика щодо академічної доброчесності регламентується положенням "Положення про систему запобігання та виявлення плагіату у Національному технічному університеті "Дніпровська політехніка". https://www.nmu.org.ua/ua/content/activity/us_documents.pdf .

У разі порушення здобувачем вищої освіти академічної доброчесності (списування, плагіат, фабрикація), робота оцінюється незадовільно та має бути виконана повторно. При цьому викладач залишає за собою право змінити тему завдання.

7.2. Комунікаційна політика

Здобувачі вищої освіти повинні мати активовану університетську пошту.

Усі письмові запитання до викладачів стосовно курсу мають надсилатися на університетську електронну пошту.

7.3. Політика щодо перескладання

Роботи, які здаються із порушенням термінів без поважних причин оцінюються на нижчу оцінку. Перескладання відбувається із дозволу деканату за наявності поважних причин (наприклад, лікарняний).

7.4. Відвідування занять

Для здобувачів вищої освіти денної форми відвідування занять є обов’язковим. Поважними причинами для неявки на заняття є хвороба, участь в університетських заходах, академічна мобільність, які необхідно підтверджувати документами. Про відсутність на занятті та причини відсутності здобувач вищої освіти має повідомити викладача або особисто, або через старосту.

За об’єктивних причин (наприклад, міжнародна мобільність) навчання може відбуватись в он-лайн формі за погодженням з керівником курсу.

7.5. Політика щодо оскарження оцінювання

Якщо здобувач вищої освіти не згоден з оцінюванням його знань він може опротестувати виставлену викладачем оцінку у встановленому порядку.

7.6. Студентоцентрований підхід

Для врахування інтересів та потреб студентів на початку вивчення курсу здобувачам вищої освіти пропонується відповісти у системі Moodle на низку питань щодо інформаційного наповнення курсу. Відповідно до результатів опитування формується траєкторія навчання з урахуванням потреб студентів.

Під час навчання здобувачі вищої освіти реалізують своє право вибору індивідуальних завдань лабораторних робіт.

Наприкінці вивчення курсу та перед початком сесії здобувачам вищої освітим пропонується анонімно заповнити у системі Moodle або Teams електронні анкети для оцінки рівня задоволеності методами навчання і викладання та врахування пропозицій стосовно покращення змісту навчальної дисципліни. За результатами опитування вносяться відповідні корективи у робочу програму та силабус.

8. Рекомендовані джерела інформації

1. Zgurovsky M.Z. Big Data: Conceptual Analysis and Applications. [Текст] / M.Z. Zgurovsky, Y.P. Zaychenko // Springer, 2021, 298 p.
2. Stanislaw Osowski. Sieci Neuronowe do Przetwarzania Informacji [Текст] / Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2000, 344 c.
3. Силен Д. Основы Data Science и Big Data. Python и наука о данных [Текст] /Д. Силен, А. Мейсман, М. Али // СПб.: Питер, 2017, 336 с.
4. Akerkar R. Models of Computation for Big Data [Текст] / R. Akerkar // Springer, 2018, 110 p.
5. Ghavami P. Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics [Текст] / P. Ghavami // CreateSpace Independent Publishing Platform, 2016, 204 p.
6. Feeney K. Engineering Agile Big-Data Systems [Текст] / K. Feeney, J. Davies, J. Welch, S. Hellmann, C. Dirschl, A. Koller, P. Francois, A. Marciniak // River Publishers, 2018, 436 p.
7. Мороз Б.І. Лабораторний практикум з курсу: “Аналіз даних та процесів”. [Электрон. ресурс]. Режим доступа: https://do.nmu.org.ua/ course/view.php?id=3214 (дата звернення: 20.08.2021).
8. Big Data Fundamentals courses [Электрон. ресурс]. Режим доступа: https://cognitiveclass.ai/learn/big-data (дата звернення: 20.08.2021).
9. Big Data Analytics [Электрон. ресурс]. Режим доступа: https://cognitiveclass.ai/learn/analytics/ (дата звернення: 20.08.2021).
10. Волкова С. Просто Big Data [Текст] / С. Волкова // СПб.: Страта, 2019, 148 с.
11. Форман Дж. Много цифр: Анализ больших данных при помощи Excel [Текст] / Дж. Форман // М.: Альпина Паблишер, 2016, 464 с.

12. Kozhevnikov A. V. Estimation of the population density spatial distribution using clutter model [Текст] / A. V. Kozhevnikov, Ye. A. Krivosheyev // Науковий вісник НГУ – Дніпропетровськ: НГУ, 2011, №2, с. 31 – 36.
13. UCI Machine Learning Repository [Электрон. ресурс]. Режим доступа: http://archive.ics.uci.edu/ml/index.php/ (дата звернення: 20.08.2021).