СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
«Аналіз даних та знань»

Ступінь освіти	Бакалавр
Галузь знань	12 Інформаційні технології
Тривалість викладання	1 семестр
Заняття	Осінній семестр
лекції	2 год./тижд.
практичні роботи	1 год./тижд.
Мова викладання	українська

Передумови для вивчення: вивчення дисципліни «Аналіз даних та знань» у встановлених відповідною робочою програмою обсягах передбачає розуміння основ програмування.

Сторінка курсу в СДО НТУ «ДП»:
https://do.nmu.org.ua/course/view.php?id=6005
Консультації: за окремим розкладом, що попередньо погоджений зі здобувачами освіти.
Онлайн-консультації: MS Teams, електронна пошта**.**

Інформація про викладача:

Викладач:

Хабарлак Костянтин Сергійович

Ph.D., асистент каф. САУ

Посилання на профілі:

Сторінка кафедри САУ

Orcid ID

Scopus ID

Google Scholar

1. Анотація курсу

Щодня людина та створені нею пристрої генерують неосяжний об’єм даних. Необроблені дані мають невисоку цінність, адже їх складно зберігати, структурувати та аналізувати. Тому все більший розвиток отримують підходи **Аналізу даних та знань.** В даному курсі здобувач познайомиться як за допомогою мови програмування Python робити статистичні оцінки даних, визначати найкращий напрямок подальшого розвитку продуктів компанії за допомогою A/B тестів, навчати моделі машинного навчання для вирішення задач регресії, класифікації та зменшення розмірності даних.

2. Мета та завдання навчальної дисципліни

Мета дисципліни – сформувати у бакалаврів навички обробки та аналізу даних за допомогою мови програмування Python та сучасних бібліотек машинного навчання та обробки даних: scikit-learn, SciPy, Pandas, NumPy, Matplotlib, що буде корисним для подальшого працевлаштування здобувача.

Завдання курсу:

- навчитися розраховувати статистичні оцінки та візуалізувати дані за допомогою мови програмування Python;
- навчитись застосовувати машинне навчання до практичних задач;
- отримати практичні навички проведення A/B тестування;
- опанувати роботу із бібліотекою машинного навчання scikit-learn (sklearn)

3. Результати навчання

1. Знати як працювати та візуалізувати дані з електронних таблиць за допомогою мови програмування Python та бібліотеки Pandas.
2. Отримати навички з роботою бібліотеки scikit-learn (sklearn) для розв'язання задач методами машинного навчання.
3. Володіти навичками проведення A/B тестування для визначення подальшого напрямку розвитку програмного продукту компанії із перевіркою статистичної значущості результатів.
4. Вміти будувати регресію та класифікувати дані. Оцінювати результати на тестовій вибірці або за допомогою крос-валідації.
5. Отримати базові навички обробки зображень і текстів за допомогою вивчених методів.

4. Структура курсу

Види та тематика навчальних занять	Обсяг складових, *години*
ЛЕКЦІЇ	**78**
1. Первинний аналіз даних Знайомство із мовою програмування Python для аналізу даних. Використання бібліотеки Pandas для табличних даних та розрахунку статистичних оцінок. Типи даних. Бібліотека SciPy.	6
2. Візуалізація даних Основні види графіків, діаграм. Їх побудова, доречність використання для наявних даних, особливості.	6
3. Бутстрап Поняття популяції та вибірки даних, зміщеної вибірки, повторного та статифікованого відбору. Алгоритм бутстрапу, розрахунок бутстрапівських оцінок.	8
4. Побудова та підгонка лінійної регресії Знайомство із бібліотекою sklearn. Інтерфейс бібліотеки. Побудова лінійної регресії за даними. Створення та використання нелінійних ознак.	8
5. Передбачення за допомогою регресії. Перехресний контроль Проблеми передбачення. Оцінка довірчого інтервалу за допомогою бутстрапу. Проведення експериментів. Поняття тренувальної та тестової вибірок, перехресного контролю (крос-валідації)	8
6. A/B тестування Знайомство із поняттям A/B тестування. Коли доречно та як необхідно проводити A/B тестування. Вклад випадковості в результати тестування. Перестановочних тест. Перевірка статистичної значущості	8
7. Дисперсійний аналіз та багаторукий бандит Проблема множинного тестування в A/B тесті. Дисперсійний аналіз. Вибір найкращого варіанта за допомогою алгоритму багаторукого бандита.	8

8. Лінійні, метричні та ймовірнісні методи класифікації Поняття задачі класифікації. Основні ідеї та методи: логістична регресія, К найближчих сусідів, наївний Баєс	10
9. Методи зменшення розмірності даних Проблема обробки даних із великою розмірністю та методи її зменшення. Метод головних компонент.	8
10. Основи аналізу текстів та зображень випадковим лісом та градієнтним бустінгом Як працювати із зображеннями та текстом за допомогою розглянутих методів. Бустінг як основа методів класифікації випадкового лісу та градієнтного бустінгу	8
ПРАКТИЧНІ ЗАНЯТТЯ	**36**
Лабораторна робота №1: Первинний аналіз даних Мета: закріпити теоретичні знання і розвинути практичні навички роботи з таблицями в Pandas, розрахунку статистик та візуалізації даних.	6
Лабораторна робота №2: Регресійний аналіз Мета: закріплення навичок збору даних, роботи із бібліотекою sklearn для побудови лінійної регресії, створення нелінійних ознак. Застосування на практиці підходу пересхесної-валідації.	6
Лабораторна робота №3: Класифікація за допомогою логістичної регресії та наївного Баєса Мета: закріплення навичок роботи із категорійними даними, їх попередньої обробки та класифікація даних.	8
Лабораторна робота №4: A/B тест та багаторукі бандити Мета: проведення власного експерименту щодо оцінки різного дизайну веб-сторінок, закріплення навичок проведення A/B тесту, алгоритму ϵ-жадібного бандиту для довільної кількості варіантів дизайну веб-сторінки та підтвердження статистичної значущості результатів експерименту.	8
Лабораторна робота №5: Основи аналізу текстів та зображень Мета: закріпити навички класифікації текстів та зображень за допомогою методів випадкового лісу та градієнтного бустінгу.	8
КОНТРОЛЬНІ ЗАХОДИ	**6**
РАЗОМ	**120**

5. Технічне обладнання та/або програмне забезпечення

Технічні засоби навчання: мультимедійні та комп'ютерні пристрої.

Засоби дистанційної освіти: Moodle, MS Teams.

Пакети прикладних програм: Python, бібліотеки: scikit-learn, SciPy Pandas, NumPy, Matplotlib (безкоштовні).

6. Система оцінювання та вимоги

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

Рейтингова шкала	Інституційна шкала
90 – 100	відмінно
75-89	добре
60-74	задовільно
0-59	незадовільно

6.2. Здобувач ступеня освіти «Бакалавр» 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 успішності за теоретичну частину курсу (максимум – 36 балів) та оцінок за виконання практичних робіт (максимум 8 балів за кожну роботу та максимальною сумарною оцінкою за всі роботи – 64 бали). Отримані бали за теоретичну частину курсу та практичні роботи додаються і є підсумковою оцінкою за вивчення навчальної дисципліни. Максимально за поточною успішністю здобувач вищої освіти може набрати 100 балів.

Шкала оцінювання (зазначено максимально можливі бали):

Теоретична частина	**Практичні роботи**		**Разом**
	При своєчасному складанні	При несвоєчасному складанні	
40	60	40	100

6.3 Критерії оцінювання поточного та підсумкового контролю:

– підсумкове оцінювання відбувається у формі диференційованого заліку у форматі тесту, який складається з 16 завдань (15 запитань із вибором варіанту відповіді – 2 бали за правильну відповідь; 1 завдання у формі задачі – максимум 6 балів, якщо надано повністю правильну і обґрунтовану відповідь);
– поточне оцінювання практичних робіт відбувається шляхом захисту звіту з відповідної роботи (максимальний бал – 8, який формується наступним чином: 50 % – правильність і повнота викладення матеріалу в звіті, 50 % – захист індивідуальної роботи шляхом відповіді на контрольні питання).

7. Політика курсу

7.1. Політика щодо академічної доброчесності. Академічна доброчесність студентів 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). У НТУ «Дніпровська політехніка» політика щодо академічної доброчесності регламентується положенням "Положення про систему запобігання та виявлення плагіату у Національному технічному університеті "Дніпровська політехніка".

У разі порушення студентом 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. Студенти повинні мати активовану університетську (корпоративну на домені @nmu.one) 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 студентів денної форми відвідування занять є обов'язковим. Поважними причинами для неявки на заняття є хвороба, участь в університетських заходах, відрядження, які необхідно підтверджувати документами у разі тривалої (два тижні) відсутності. Про відсутність на занятті та причини відсутності студент має повідомити викладача або особисто, або через старосту. Якщо студент захворів, ми рекомендуємо залишатися вдома і навчатися за допомогою дистанційної платформи. Студентам, чий стан здоров'я є незадовільним і може вплинути на здоров'я

інших студентів, буде пропонуватися залишити заняття (така відсутність вважатиметься пропуском з причини хвороби). Лабораторні заняття не проводяться повторно, ці оцінки неможливо отримати під час консультації. **За об'єктивних причин (наприклад, міжнародна мобільність) навчання може відбуватись в он-лайн формі за погодженням з керівником курсу.**

7.5. Участь в анкетуванні. Наприкінці вивчення курсу та перед початком сесії студентам буде запропоновано анонімно заповнити електронні анкети (MS Office 365), які буде розіслано на ваші університетські поштові скриньки. Заповнення анкет є важливою складовою вашої навчальної активності, що дозволить оцінити дієвість застосованих методів викладання та врахувати ваші пропозиції стосовно покращення змісту навчальної дисципліни.

8. Рекомендовані джерела інформації

1. Математичні методи інтелектуального аналізу даних: [навчальний посібник для здобувачів першого рівня вищої освіти спеціальності 124 Системний аналіз] / Т. Шабельник, О. Дяченко. – Маріуполь: МДУ, 2021. – 163 с
2. Кононова К. Ю. Машинне навчання: методи та моделі / К. Ю. Кононова. – Харків: ХНУ імені В. Н. Каразіна, 2020. – 301 с.
3. Документація бібліотеки машинного навчання scikit-learn. URL: https://scikit-learn.org (дата звернення: 02.11.2023).
4. Документація бібліотеки аналізу даних в Python: pandas. URL: https://pandas.pydata.org/ (дата звернення: 02.11.2023).
5. Practical Statistics for Data Scientists / P. Bruce, A. Bruce, P. Gedeck. – O'Reilly Media, 2020.