



Національний технічний
університет України «КИЇВСЬКИЙ
ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені
ІГОРЯ СІКОРСЬКОГО»



Кафедра
обчислювальної
техніки



Sigma Software
University

**Освітній компонент сертифікатної програми
«Технології аналізу та обробки даних (Data Science) в комп'ютерних
системах»**

**ОСНОВИ НАУКИ ПРО ДАНІ
Робоча програма навчальної дисципліни (Силабус)**

Реквізити навчальної дисципліни

Рівень вищої освіти	<i>Перший (бакалаврський)</i>
Галузь знань	<i>12 Інформаційні технології</i>
Спеціальність	<i>121 Інженерія програмного забезпечення; 123 Комп'ютерна інженерія, 126 Інформаційні системи та технології</i>
Освітня програма	<i>Комп'ютерні системи та мережі Інженерія програмного забезпечення комп'ютерних систем Інтегровані інформаційні системи Інформаційні управляючі системи та технології Інформаційне забезпечення робототехнічних систем Інженерія програмного забезпечення інформаційних систем</i>
Статус дисципліни	<i>Вибіркова</i>
Форма навчання	<i>Очна (денна)</i>
Рік підготовки, семестр	<i>3 курс, 5 семестр</i>
Обсяг дисципліни	<i>4 кредити/120 год. денна форма: лекцій 36 годин, Пр. роб. 18 год., СРС 66 год.</i>
Семестровий контроль/ контрольні заходи	<i>Залік</i>
Розклад занять	<i>Згідно розкладу на 5-й семестр поточного навчального року за адресою https://schedule.kpi.ua</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	<i>Лектор: д.т.н, професор, Новотарський Михайло Анатолійович novotar@gmail.com, https://bit.ly/3XAqgCq Лабораторні: д.т.н, професор, Новотарський Михайло Анатолійович novotariskyi.mykhailo@iit.kpi.ua</i>
Розміщення курсу	<i>Навчальний курс розміщений на платформі дистанційного навчання «Сікорський» в середовищі Google Workspace for Education: https://classroom.google.com/c/NDU4MDUzNDQ0MDU4?cjc=Iwqulmf</i>

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Метою вивчення курсу «Основи науки про дані» є вивчення сучасних методів та технологій добування даних, підготовка студентів у виборі та використанні методів початкової підготовки, очистки та побудови моделей.

Навчальна дисципліна «Основи науки про дані» призначена для надання базових знань у галузі аналізу даних для розв'язування наукових та інженерно-технічних задач та програмування з врахуванням особливостей сучасних підходів та методів добування даних.

Навчальна дисципліна «Основи науки про дані» є вибірковою дисципліною.

В рамках даного курсу розглядається основне поняття про дані та їх властивості, обговорюються основні різновидності даних та їх атрибутів. Коротко наводиться інформація про способи представлення даних.

Далі розглядається проблема аналізу даних для отримання корисних знань. Розглядаються основні етапи аналізу даних та методи, які в даному випадку використовуються. Детально розглянуто такі задачі добування даних, як класифікацію та кластеризацію, а також задачі прогнозування та візуалізації.

Курс «Основи науки про дані» включає ознайомлення студентів з основними етапами добування даних та їх реалізацією. На початковому етапі розглянуто принципи початкової підготовки даних та побудови моделі, а на наступних етапах – принципи очистки даних та аналізу отриманих результатів.

Предметом дисципліни є:

- методи аналізу даних;
- методи очистки даних;
- методи побудови моделей.

Студент буде знати:

основні принципи розробки алгоритмів та програмного забезпечення розв'язування задач добування даних на ПК;

досліджувати алгоритми добування даних,

виявляти їх переваги та недоліки,

обирати оптимальні алгоритми розв'язування задач, обробки даних та розробляти програми розв'язування задач;

виконувати аналіз і опрацювання результатів розв'язування задач,

використовувати методи оптимізації.

Завданням дисципліни є формування і підсилення таких компетентностей:

Для спеціальності 121 Інженерія програмного забезпечення

- Здатність до абстрактного мислення, аналізу та синтезу.
- Здатність застосовувати знання у практичних ситуаціях.
- Здатність вчитися і оволодівати сучасними знаннями.
- Здатність ідентифікувати, класифікувати та формулювати вимоги до програмного забезпечення.
- Здатність формулювати та забезпечувати вимоги щодо якості програмного забезпечення у відповідності з вимогами замовника, технічним завданням та стандартами.
- Здатність накопичувати, обробляти та систематизувати професійні знання щодо створення і супроводження програмного забезпечення та визнання важливості навчання протягом всього життя.
- Здатність до алгоритмічного та логічного мислення.

Для спеціальності 123 «Комп'ютерна інженерія»

- Здатність вчитися і оволодівати сучасними знаннями.
- Здатність працювати в команді.
- Здатність використовувати сучасні методи і мови програмування для розроблення алгоритмічного та програмного забезпечення.
- Здатність проектувати системи та їхні компоненти з урахуванням усіх аспектів їх життєвого циклу та поставленої задачі, включаючи створення, налаштування, експлуатацію, технічне обслуговування та утилізацію.
- Здатність розробляти, адаптувати, використовувати програмно-технічні засоби для збирання, оброблення та інтелектуального аналізу даних в комп'ютерних системах: паралельних, розподілених, хмарних, безпечних, інтелектуальних, розумних, Інтернет речей, зокрема, з використанням штучного інтелекту.

Для спеціальності 126 Інформаційні системи та технології

- Здатність до абстрактного мислення, аналізу та синтезу.
- Здатність аналізувати об'єкт проектування або функціонування та його предметну область.
- Здатність застосовувати стандарти в області інформаційних систем та технологій при розробці функціональних профілів, побудові та інтеграції систем, продуктів, сервісів і елементів інфраструктури організації.
- Здатність до проектування, розробки, налагодження та вдосконалення системного, комунікаційного та програмно апаратного забезпечення інформаційних систем та технологій.
- Здатність управляти та користуватися сучасними інформаційно-комунікаційними системами та технологіями (у тому числі такими, що базуються на використанні Інтернет).
- Здатність розробляти структури робототехнічних систем та проектувати функціональні пристрої роботів та робототехнологічних комплексів.
- Здатність формувати нові конкурентоспроможні ідеї й реалізовувати їх у проектах (стартапах)

Як додатковий результат навчання вибірково навчальна дисципліна «Основи науки про дані» спрямована на формування здатності використовувати прийоми ефективного вибору метрик для оцінки точності моделей, розробляти спеціалізоване програмне забезпечення з використанням сучасних бібліотек для машинного навчання та аналізу даних.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Матеріал навчальної дисципліни пов'язаний з матеріалами, які вивчалися у курсах «Вища математика: Диференціальні обчислення, лінійна алгебра», «Програмування» та «Дискретна математика».

Знання та практичні навички, які отримані в рамках даної навчальної дисципліни, можуть бути застосовані при вивченні наступних курсів: «Комп'ютерне моделювання», «Основи програмної інженерії», «Системне програмування», «Технологія розподілених обчислень», «Мережні та інформаційні технології» та інші.

3. Зміст навчальної дисципліни

Навчальна дисципліна «Основи науки про дані» включає вивчення наступних тем.

Тема 1. Вступ. Базові відомості про дані. Поняття даних. Набір даних та їх атрибутів. Вимірювання. Шкали. Типи наборів даних: записи, хімічні дані, графи. Формати зберігання даних. Основні положення про бази даних. Системи управління базами даних. Класифікація видів даних. Метадані.

Тема 2. Основні положення аналізу даних. Поняття про процес добування даних, процес KDD (Knowledge Discovery in Databases). Математична база методів добування даних, статистичні методи. Причини виникнення та мотивація розвитку методів добування даних. Суть технології добування даних, мультидисциплінарність. Порівняння методів машинного навчання та методів добування даних. Проблеми підготовки даних. Огляд алгоритмів та методів попередньої обробки даних: очистка, перетворення та скорочення.

Тема 3. Базові поняття про бібліотеку Pandas. Структура даних Pandas Series. Основні характеристики бібліотеки. Особливості інсталяції під різними операційними системами. Переваги структури Series у порівнянні з масивами NumPy. Визначення масиву міток за замовчуванням та у вигляді рядків. Способи виводу даних типу Series. Доступ до елементів Series та їх модифікація. Створення Series з використанням масиву NumPy. Динамічне копіювання посилань на елементи при створенні Series. Створення Series з словника. Фільтрація елементів Series. Групові математичні операції Series з скалярними величинами. Операції /, *, -, + з двома об'єктами Series. Робота з повторюваними елементами в Series. Використання NaN: перевірка наявності значень NaN, фільтрація елементів з NaN

Тема 4. Основні характеристики та способи застосування структури даних DataFrame. Переваги структури DataFrame у порівнянні з двовимірними масивами NumPy. Створення DataFrame за замовчуванням та з використанням словника. Створення міток рядків у DataFrame.

Генерування DataFrame з використанням методу `reshape()`. Перетворення DataFrame у NumPy масив. Зчитування стовпця DataFrame у Series. Зчитування рядків DataFrame. Додавання стовпця з однаковими значеннями. Оновлення значень у стовпці. Модифікація значення елемента та перевірка його входження. Фільтрація у структурі DataFrame. Повторення даних. Арифметичні методи `add()`, `sub()`, `div()`, `mul()`. Арифметичні операції між DataFrame і Series. Проблема відображення даних DataFrame: поворот структури даних та приклади його застосування. Поворот з одним та кількома стовпцями. Типові помилки при використанні методу повороту. Методи `stack()` і `unstack()` та приклади їх застосування. Застосування функції `melt()`, пропуск стовпців при використанні `melt()`.

Тема 5. Функції бібліотеки Pandas. Модифікація великих наборів даних. Функції округлення: `trunc()`, `around()`, `floor()`, `ceil()`. Функції взяття логарифмів: `log2()`, `log10()`, `log()`. Функції обчислення суми: `sum()`, `cumsum()`. Функції обчислення добутку: `prod()`, `cumprod()`. Функція обчислення скінченних різниць: `diff()`. Знаходження найменшого спільного кратного (LCM - Lowest Common Multiple). Знаходження найбільшого спільника дільника (GCD - Greatest Common Denominator). Тригонометричні функції та функції множин. Функції для стовпців. Функція статистики. Сортвання Series. Сортвання DataFrame. Кореляція і коваріантність у Series та DataFrame. Інструменти I/O у Pandas: робота з файлами різних типів. Перегляд фрагментів набору даних. Визначення типів даних колонок. Роздруківка технічної інформації про набір даних. Визначення кількості NaN елементів у наборі даних. Заміна відсутніх даних константним значенням. Заповнення середнім значенням по стовпцю. Заповнення інтерпольованими даними. Видалення рядків та стовпців з пропусками. Встановлення порогу пропусків. Розбиття на відрізки однакового розміру. Злиття двох структур DataFrame.

Тема 6. Індексація та реіндексація наборів даних. Вирівнювання даних у Pandas. Зміна порядку виводу наявних даних з маркером відсутності. Метод `reindex()` для Series та для DataFrame. Реіндексація з ключовим словом `axis`. Використання індексації стороннього набору даних. Метод `align()`. Вирівнювання об'єктів з використанням `align(s, join="outer")`, `align(s, join="inner")`, `align(s, join="left")`, `align(s, join="right")`. Метод переіменування `rename(str.upper)`. Метод переіменування `rename()` з відображенням. Динамічне перейменування по осях.

Тема 7. Багаторівнева індексація. Створення та використання об'єкта Multiindex.

Створення об'єкта `MultiIndex` з використанням методів `from_tuples()`, `from_product()`, `from_frame()`. Автоматичний `MultiIndex`. Керування відображенням індексів набору даних. Використання кортежів, як міток. Вирівнювання даних з `MultiIndex`. Застосування `reindex()` для Series та DataFrame. Розширене індексування з ієрархічним індексом з методом `loc()`. Використання зрізів. Перестановка рівнів за допомогою `swaplevel`. Зміна порядку рівнів за допомогою `reorder_levels`. Метод `rename()` для перейменування `MultiIndex`. Перейменування основного індексу. Застосування методу `MultiIndex.set_names()`. Сортвання `MultiIndex`. Сортвання `MultiIndex` по рівнях.

Задавання `level` по номеру для параметру `index` та `columns`. Метод `is_monotonic_increasing()`. Зріз по сортованому `MultiIndex`. Метод `take()` для структур Series та DataFrame. Індексація за допомогою `IntervalIndex`. Дані типу `Categoricals`. Перетворення Series у тип `category`. Бінінг (створення контейнерів) даних за допомогою `cut`. Створення діапазонів інтервалів. Застосування діапазонів як `IntervalIndex` у DataFrame. Діапазони з параметром `freq`. Агрегування даних по єдиному стовпцю. Застосування методів агрегування у параметрі `aggfunc`. Метод множинної агрегації в `pandas DataFrame`. Використання різних агрегацій по стовпцях. Сортвання зведеної таблиці

Тема 8. Відображення великих наборів даних. Бібліотека Matplotlib. Способи інсталяції Matplotlib. Відображення найпростішого графіка. Графік з кумулятивною сумою. Відображення DataFrame з використанням міток. Функціональна залежність одного стовпця від іншого. Графік типу `Bar` для структур Series та DataFrame. Графік типу «складений `Bar`» вертикальний. Графік типу «складений `Bar`» горизонтальний. Гістограма для структури DataFrame. Складена гістограма. Горизонтальна кумулятивна гістограма. Групові гістограми по стовпцях. Діаграми площ у стопку. Діаграми площ, які накладаються. Діаграма розсіювання (точкова діаграма). Багатоколірна мультидіаграма розсіювання по стовпцях. Кругова діаграма для DataFrame. Модифікація кругових діаграм. Матрична діаграма розсіювання. Графік щільності. Криві Ендрюса. Графік відставання.

Графік автокореляції. Бутстреп графік. RadViz. Загальні аргументи стилю графіка. Керування легендою. Керування мітками. Логарифмічні шкали. Побудова графіка на вторинній осі ординат. Графік таблиця. Роздруківка списку кольорових стилів. Процес формування графіка з використанням стилю. Темний фон на графіку. Створення кількох графіків методом subplots(). Використання спільних осей Y та X.

Тема 9. Задача логістичної регресії з використанням бібліотеки PyTorch. Інсталяція PyTorch під Anzconda. Перевірка працездатності бібліотеки PyTorch. Основи положення роботи з PyTorch. Створення тензора з нульовим або одиничним значенням. Створення довільного тензора. Перетворення pandas Series у PyTorch tensor. Перетворення pandas DataFrame у PyTorch tensor. Стандартні набори даних. Набор даних MNIST. Завантаження MNIST. Перетворення формату рисунка. Підготовка даних до навчання за допомогою DataLoaders. Теоретичні основи логістичної регресії. Прямий хід логістичної регресії. Початкова ініціалізація параметрів моделі. Обчислення у з урахуванням градієнтного спуску для параметрів. Застосування softmax(). Два варіанти обчислення softmax(). Крос-ентропійна функція втрат. Два варіанти обчислення крос-ентропійної втрати. Зворотний хід логістичної регресії. Тренування моделі. Перевірка точності. API вищого рівня логістичної моделі. Об'єктно-орієнтована рефакторизація. Використання torch.nn. Використання torch.nn.Module.

Тема 10. Створення глибокої нейронної мережі з використанням PyTorch. Навчальний набір Fashion MNIST. Код завантаження FashionMNIST. Ітерування і візуалізація набору даних Fashion-MNIST. Створення власного набору даних Custom Dataset. Методи __init__, __len__, __getitem__. Код підготовки даних у DataLoader. Ітерування у DataLoader. Перетворення даних. Побудова нейронної мережі. Визначення класу NeuralNetwork. Використання моделі. Шари моделі. Диференціювання з використанням torch.autograd. Тензори, функції та обчислювальний графік. Обчислювальні градієнти. Вимкнення відстеження градієнта. Оптимізація модельних параметрів. Гіперпараметри. Цикл оптимізації. Функція втрат. Способи оптимізації моделі.

4. Навчальні матеріали і ресурси

4.1. Базова література:

1. Основи науки про дані [Електронний ресурс]: навч. посіб. для студ. спеціальності 123,121 / М. А. Новотарський; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 5,5 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 294 с.
2. Новотарський М.А. Лекції х курсу «Основи науки продані» // <https://cloud.comsys.kpi.ua/s/pjw4nAStZFPsDos>
3. Новотарський М.А. Методичні вказівки до виконання лабораторних робіт з курсу «Основи науки продані» // Новотарський М.А. Лекції з курсу «Основи науки продані» // <https://classroom.google.com/c/NDU4MDUzNDQ0MDU4?cjc=lwqulmf>
4. Системи Data Science. Практикум з системної інженерії систем Data Science [Електронний ресурс] / КПІ ім. Ігоря Сікорського ; уклад. П. П. Масляноко. – Електронні текстові дані (1 файл: 2,3 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 68 с.
5. Новотарський М.А., Нестеренко Б.Б. Штучні нейронні мережі: обчислення.– К.: Інститут математики НАН України, 2004.– 408 с.
6. Нестеренко Б.Б., Новотарський М.А. Методи розв'язування крайових задач на нейронних мережах // Збірник праць Інституту математики НАН України.– 2006.– Т.3, №4.– С.231-238.
7. Нестеренко Б.Б., Новотарський М.А. Формальні засоби моделювання паралельних процесів та систем // Праці Інституту математики НАН України.– Т.90.–Київ: Ін-т математики НАН України, 2012.–334с.

4.2. Додаткова література:

1. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. – К.: ІСЗІ КПІ ім. Ігоря Сікорського», 2018. — 297с.
2. Гороховатський, Володимир Олексійович. Методи інтелектуального аналізу та оброблення даних: навчальний посібник / В.О. Гороховатський, І.С. Творошенко ; Міністерство освіти і науки України, Харківський національний університет радіоелектроніки. – Харків : ХНУРЕ, 2021. – 90 с. https://opac.kpi.ua/F/?func=direct&doc_number=000634401&local_base=KPI01
3. Інтелектуальний аналіз даних : навчальний посібник / укладачі: Я.М. Дрінь, І.В. Малик, Ю.А. Літвінчук ; Міністерство освіти і науки України, Чернівецький національний університет імені Юрія Федьковича. – Чернівці : Чернівецький національний університет імені Юрія Федьковича, 2020. – 95 с. https://opac.kpi.ua/F/?func=direct&doc_number=000615829&local_base=KPI01
4. Юрченко, Ігор Валерійович. Прикладний статистичний аналіз з використанням Python : навчальний посібник / І.В. Юрченко ; Міністерство освіти і науки України, Чернівецький національний університет імені Юрія Федьковича. Видання 3-є, доповнене. – Чернівці : Чернівецький національний університет, 2021. – 107 стор. Замовити в Бібліотеці КПІ: https://opac.kpi.ua/F/?func=direct&doc_number=000635029&local_base=KPI01

4.1. Інформаційні ресурси

1. Основи науки про дані / дистанційний курс // <https://classroom.google.com/c/NDU4MDUzNDQ0MDU4?cjc=Iwqulmf>
2. <https://www.kaggle.com/>
3. <https://www.jetbrains.com/>
4. <https://github.com/PacktPublishing/Artificial-Intelligence-with-Python>
5. <https://pandas.pydata.org/>
6. <https://scapy.net/>
7. <https://developers.google.com/optimization>
8. <https://www.tensorflow.org/>
9. <https://keras.io/>
10. <https://opencv.org/>
11. <https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html>
12. <https://www.statsmodels.org/stable/examples/notebooks/generated/ols.html>
13. <https://scikit-learn.org/stable/modules/sgd.html#regression>

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Розподіл навчального часу за видами занять і завдань з дисципліни згідно з робочим навчальним планом: на кредитний модуль виділено 120 годин та 4 кредитів – 18 лекцій (36 годин), 5 лабораторних робіт (18 годин), самостійна робота (66 годин), заліковий тест.

Для досягнення мети навчальної дисципліни студенти мають отримати знання про можливі види даних та способи їх обробки, про типові задачі, які виникають при обробці даних. Особлива увага приділяється використанню сучасних інструментів попередньої підготовки даних, зокрема бібліотеці Pandas, бібліотеці Matplotlib та ряду інших бібліотек, використання яких дозволяє пройти повний цикл підготовки даних перед обробкою. Це забезпечить студентам розуміння важливості попереднього аналізу даних та знання технологій підготовки цільових наборів даних у рамках

процесу KDD. У рамках курсу також передбачено знайомство з основними принципами аналізу даних на базі бібліотек TensorFlow і PyTorch.

Лабораторний практикум.

Дані лабораторні роботи виконують шляхом модифікації початкових даних, які знаходяться у зошиті, номер якого відповідає номеру лабораторної роботи. Тому необхідно попередньо створити папку та завантажити в цю папку всі початкові зошити. Необхідно також уміти створити оригінальні зошитів jupyter notebook.

Для успішного захисту лабораторних робіт разом зі звітами викладачеві необхідно надати зошити з виконаними завданнями для кожної лабораторної роботи. Звіти та зошити jupyter notebook необхідно завантажувати у відповідне завдання, яке створене системі Google classroom. Перевірка теоретичних знань проводиться персонально виключно на заняттях з лабораторних робіт у режимі відеоконференції.

Мета лабораторних робіт — набуття практичних навичок у розробці програмних застосунків для початкової обробки даних та формування цільових наборів даних, а саме:

- Технологія вибору відповідних даних з DataFrame або Series, вибір фрагментів даних з набору даних.
- Технологія швидкого та ефективного вибору необхідних даних з набору даних з використанням структур даних та інструментів бібліотеки Pandas.
- Технологія роботи з наборами даних, які містять відсутні або помилкові дані.
- Робота з простою моделлю лінійної регресії.
- Робота з простим алгоритмом кластеризації, який відносять до алгоритмів аналізу ринкового кошику.

Практична складова включає захист лабораторних робіт, перелік яких наведено в таблиці 1.

Таблиця 1

Перелік лабораторних робіт

№ з/п	Назва лабораторної роботи (комп'ютерного практикуму)	Кількість ауд. годин
1	Лабораторна робота № 1. Підсумкові функції та відображення Теоретична складова: Тема 2,Тема 3.	3
2	Лабораторна робота № 2. Індексуння, вибір, редагування набору даних. Теоретична складова: Тема 4, Тема 5.	3
3	Лабораторна робота № 3. Робота з пропущеними даними Теоретична складова: Тема 6, Тема 7,Тема 8.	5
4	Лабораторна робота № 4. Розв'язування задачі лінійної регресії Теоретична складова: Тема 9	4
5	Лабораторна робота №5. Розв'язування задачі кластеризації Теоретична складова: Тема 10	3
Всього		18 годин

6. Самостійна робота здобувача вищої освіти денної форми навчання

Види самостійної роботи (66 години):

- підготовка до аудиторних занять
- підготовка лабораторних робіт

Самостійна робота студента складається з теоретичної та практичної складової. Теоретична складова передбачає вивчення додаткового матеріалу, що поглиблює знання, які отримані на лекції. Матеріал, який необхідно додатково вивчити, та літературні джерела для вивчення даного матеріалу вказані у таблиці 2. Всього 20 годин.

Основне завдання циклу лабораторних занять полягає у набутті студентами необхідних практичних навичок розробки алгоритмів та програмного забезпечення для розв'язування задач обробки даних на ПК. Всього 46 годин.

Таблиця 2

Тематика самостійної роботи студента

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)	Кількість годин на СРС
1	Вступ. Базові відомості про дані. Види даних та способи їх обробки. Операції над даними. Класифікація інформації за видами і типами. Види кодувань. Структура даних та одиниці вимірювання. Література: https://hi-news.pp.ua/kompyuteri/1379-vidi-danih-sposobi-yih-obrobki.html	2
2	Основні положення аналізу даних. Поняття про взаємозв'язок статистичних показників. Види та форми зв'язків між явищами. Прийоми виявлення щільності зв'язку між показниками досліджуваних явищ. Література: https://xreferat.com/22/2260-1-osnovn-polozhennya-kompleksnogo-statistichnogo-anal-zu-danih-u-pravov-iy-statistic.html	2
3	Базові поняття про бібліотеку Pandas. Загальні поняття про Pandas. Сфера застосувань бібліотеки Pandas. Додаткові властивості структур даних у Pandas/ Література: https://coderlessons.com/tutorials/mashinnoe-obuchenie/uchebnik-tensorflow/6-uchebnik-po-python-pandas	2
4	Основні характеристики та способи застосування структури даних DataFrame. class pandas.DataFrame. Атрибути та методи структури даних DataFrame Література: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html	2
5	Функції бібліотеки Pandas. Модифікація великих наборів даних. Бінарні функції. Функції порівняння значень. Перевірка на еквівалентність. Порівняння масивів. Об'єднання даних, які перекриваються. Методи описової статистики. Підрахунок значень та мода значень. Дискретизація неперервних значень. Література: https://pyprog.pro/pd/pd_11_basic_functionality.html	2
6	Індексація та реіндексація наборів даних. Різні варіанти індексації. Методи для індексації. Література: https://coderlessons.com/tutorials/python-technologies/vyuchit-python-panda/python-pandas-indeksirovanie-i-vybor-dannykh	2
7	Багаторівнева індексація. Створення та використання об'єкта Multiindex. Ієрархічне індексування. Створення об'єкта MultiIndex (ієрархічний індекс). Відновлення міток рівнів. Базове індексування на осі за допомогою MultiIndex. Визначені рівні. Розширене індексування з ієрархічним індексом. Використання слайсерів. Поперечний переріз. Література https://pandas.pydata.org/docs/user_guide/advanced.html	2
8	Відображення великих наборів даних. Бібліотека Matplotlib. Оформлення графіків. Розташування графіків. Робота з осями. Інтерактивні можливості. Література: https://jenyay.net/Matplotlib/Matplotlib	2

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)	Кількість годин на СРС
9	Задача логістичної регресії з використанням бібліотеки PyTorch. Установка PyTorch. Підготовка даних в PyTorch. Навчання моделі за допомогою PyTorch. Розгортання моделі за допомогою Windows. Література: http://cloud-5.bitp.kiev.ua/?page_id=605 https://docs.microsoft.com/ru-ru/windows/ai/windows-ml/tutorials/pytorch-installation	2
10	Створення глибокої нейронної мережі з використанням PyTorch. Класифікація зображень за допомогою PyTorch. Проблема класифікації. Завантаження даних PyTorch. Створення навчального набору даних. Волідація та контрольні набори даних. Література: https://habr.com/ru/company/piter/blog/512158/	2

Всього 20 годин

Кількість годин на самостійну підготовку лабораторних робіт для кожного студента наведено в таблиці 3.

Таблиця 3.

Лабораторні роботи

№ з/п	Назва лабораторної роботи	Кількість годин на самостійну підготовку
1	Лабораторна робота № 1. Підсумкові функції та відображення.	9
2	Лабораторна робота № 2. Індексвання, вибір, редагування набору даних.	9
3	Лабораторна робота № 3. Робота з пропущеними даним	9
4	Лабораторна робота № 4. Розв'язування задачі лінійної регресії	10
5	Лабораторна робота №5. Розв'язування задачі кластеризації	9

Всього 46 годин

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Під час занять з навчальної дисципліни «Основи науки про дані» студенти повинні дотримуватись певних дисциплінарних правил:

- забороняється запізнюватись на заняття;
- при вході викладача, на знак привітання, особи, які навчаються в КПІ ім. Ігоря Сікорського, повинні встати;
- не допускаються сторонні розмови або інший шум, що заважає проведенню занять;
- не допускається користування мобільними телефонами та іншими технічними засобами без дозволу викладача.

Лабораторні роботи здаються особисто з перевіркою теоретичних знань, які необхідні для виконання лабораторної роботи. Перевірка практичних результатів включає перевірку коду та виконання тестових завдань.

В процесі навчання викладач має право нарахувати до 5 заохочувальних балів за дострокове виконання лабораторної роботи, за проявлений творчий підхід при виконанні індивідуального

завдання або за активну участь у обговоренні питань, що пов'язані з тематикою лекції або практичного заняття.

За виконання та здачу лабораторної роботи після зазначеного дедлайну, за значну кількість пропущених занять, або за порушення правил поведінки на заняттях викладач може призначити до 5 штрафних балів. Під час військового стану штрафні бали не нараховуються.

При проведенні контрольних заходів та при виконанні лабораторних робіт студенти мають дотримуватися правил академічної доброчесності. При виявленні значного відсотку списування або плагіату викладач може відмовити у прийнятті даної роботи та вимагати доброчесного виконання навчального плану.

8. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

Види контролю з навчальної дисципліни «Основи науки про дані» включають:

Лабораторні роботи:

Заплановано самостійне виконання 5 лабораторних робіт.

Теми лабораторних робіт узгоджені у часі та за змістом з темами лекцій. Виконання лабораторних робіт у повному обсязі дозволяє набути практичних навичок застосування технологій початкової обробки даних.

Поточний контроль:

Передбачено 2 календарних тестування закритими тестами у системі TCXAM, які повністю охоплюють тематику даної навчальної дисципліни. Кожний поточний закритий тест містить 12 питань та триває 15 хв. У випадку дистанційного навчання закритий поточний тест проводиться на початку лекції, яка слідує за лекцією, що завершує чергову тему. При очному навчанні час чергового поточного тестування призначається викладачем за узгодженням зі студентами.

Залік

Необхідною умовою допуску студента до заліку є його індивідуальний семестровий рейтинг не менший, ніж 30 балів, відсутність повної заборгованості з лабораторних робіт та не менше, ніж одна позитивна атестація. За невиконання хоча б однієї зі згаданих вимог студент до заліку не допускається.

Студенти можуть зараховувати виконання лабораторних робіт з даного курсу за рахунок отримання сертифікатів на навчальних платформах. Умови зарахування та кількість сертифікатів узгоджується з викладачем, але не більше, ніж 2 лабораторні роботи за один сертифікат.

Залік проводиться у вигляді співбесіди зі студентом для об'єктивного визначення рівня знань, умінь та практичних навичок, отриманих за семестр. За вибором викладача залік може проводитися у вигляді контрольного тесту у системі TCXAM та складається з 20 питань. Правильна відповідь на кожне запитання тесту додає 1 бал.

Оскільки кредитний модуль має семестрову атестацію у вигляді заліку, рейтингова система оцінювання побудована за типом PCO – 1. Рейтинг студента з кредитного модуля складається з балів, які він отримує за види робіт відповідно до таблиці 4.

Таблиця 4

Оцінювання окремих видів навчальної роботи студента

4 курс			
Вид навчальної роботи	Оцінка за теоретичні знання	Оцінка за практичне виконання завдання	Всього
Виконання та захист лабораторної роботи №1	5	5	10

Виконання та захист лабораторної роботи №2	5	5	10
Виконання та захист лабораторної роботи №3	5	5	10
Виконання та захист лабораторної роботи № 4	5	5	10
Виконання та захист лабораторної роботи № 5	5	5	10
Календарний контроль			30
Залік			20
Усього за семестр			100

Індивідуальний семестровий рейтинг (RD) студента з навчальної дисципліни складається з балів, які він отримує за:

Календарний контроль: тематичні тестування закритими тестами – 2x15=30 балів.

Залік за умови тестування закритими тестами – 20 балів,

Виконання лабораторних робіт.

Протягом семестру студенти виконують 5 лабораторних робіт.

Максимальна кількість балів за кожну лабораторну роботу – 5+5=10 балів.

Разом за лабораторні роботи (максимальна кількість балів) – 50.

Виконання залікового контрольного тестування.

В процесі навчання викладач має право нарахувати до 5 заохочувальних балів за дострокове виконання лабораторної роботи, за проявлений творчий підхід при виконанні індивідуального завдання або за активну участь у обговоренні питань, що пов'язані з тематикою лекції або практичного заняття.

Розрахунок розміру шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру становить:

$R = \sum_k r_k$, де r_k максимальний рейтинговий бал кожного з контрольних заходів (контрольне тестування, лабораторні роботи, преміальні бали, мінус штрафні бали).

Індивідуальний семестровий рейтинг студента (підсумкова семестрова рейтингова оцінка RD) є сумою балів, отриманих студентом протягом семестру за участі у передбачених контрольних заходах (контрольні та лабораторні роботи та залік).

Сума підсумкової семестрової (RD) та залікової рейтингових оцінок у балах становить підсумкову семестрову рейтингову оцінку, яка перераховується в оцінки за національною шкалою та шкалою ECTS (табл. 5).

Таблиця 5

Відповідність рейтингових балів оцінкам за університетською шкалою

Кількість балів	Оцінка
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

9. Додаткова інформація з дисципліни (освітнього компонента)

Викладання дисципліни «Основи науки про дані» для спеціальностей 121, 123 і 126 має свою специфіку, яка пов'язана з тим, що аналіз даних потребує попередньої підготовки цільових наборів даних. Для формування таких наборів використовують процес KDD, який розділяє етапи попередньої підготовки даних. В рамках даних етапів використовують спеціальні бібліотеки, технології застосування яких описуються у даному навчальному курсі.

Робочу програму навчальної дисципліни (силабус):

Складено: д.т.н. проф. Новотарський Михайло Анатолійович

Ухвалено кафедрою ОТ (протокол №12 від 23.06.2025)

Погоджено: Методичною комісією ФІОТ (протокол № 11 від 27.06.2025)