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

Кафедра цифрових технологій в енергетиці

ОБРОБКА НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***другий (магістерський)***
Галузь знань	*F Інформаційні технології*
Спеціальність	*F3 Комп'ютерні науки*
Освітня програма	*Технології візуальних обчислень*
Статус дисципліни	*Нормативна*
Форма навчання	*очна (денна)*
Рік підготовки, семестр	*1 курс осінній семестр*
Обсяг дисципліни	*На засвоєння дисципліни передбачено 120 год / 4 кредити ЄКТС: 30 год лекцій, 8 год практичних занять, 14 год лабораторних робіт, 68 год самостійної роботи.*
Семестровий контроль/ контрольні заходи	*Екзамен, МКР*
Розклад занять	*http://rozklad.kpi.ua/*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: *асистент, доктор філософії, Здор Костянтин Андрійович, email: Kostya9919moonlight@gmail.com* Лабораторні: *асистент, доктор філософії, Здор Костянтин Андрійович, email: Kostya9919moonlight@gmail.com*
Розміщення курсу	https://campus.kpi.ua/ https://colab.research.google.com/

Програма навчальної дисципліни

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

Обробка надвеликих масивів даних — це сукупність підходів, інструментів і методів, що дозволяють зберігати, обробляти та аналізувати дані, обсяг, швидкість надходження або різноманітність яких унеможливлюють застосування традиційних засобів. Дисципліна поєднує дві складові: інфраструктурну (розподілене зберігання та обробка даних, архітектури сховищ, моделі розподілених обчислень) та аналітичну (методи машинного та глибокого навчання, які застосовуються до великих обсягів даних). Отримані знання дозволяють проєктувати конвеєри обробки даних, обґрунтовано обирати модель машинного навчання під конкретну задачу та оцінювати її якість, а також враховувати обмеження, які накладає масштаб даних на вибір алгоритму та архітектури рішення.

Метою кредитного модуля є формування у студентів компетентностей у відповідності до ОПП.

ЗК 1	Здатність до абстрактного мислення, аналізу та синтезу.
ЗК 2	Здатність застосовувати знання у практичних ситуаціях.
ЗК 5	Здатність вчитися й оволодівати сучасними знаннями.
ФК 1	Усвідомлення теоретичних засад комп'ютерних наук.

ФК 3	Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області.
ФК 4	Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проєктних рішень.
ФК 5	Здатність розробляти, описувати, аналізувати та оптимізувати архітектурні рішення інформаційних та комп'ютерних систем різного призначення.
ФК 6	Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі комп'ютерних наук.
ФК 7	Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень.
ФК 8	Здатність розробляти і реалізовувати проєкти зі створення програмного забезпечення, у тому числі в непередбачуваних умовах, за нечітких вимог та необхідності застосовувати нові стратегічні підходи, використовувати програмні інструменти для організації командної роботи над проєктом.
ФК 9	Здатність розробляти та адмініструвати бази даних та знань.
ФК 15	Здатність до проектування та програмної реалізації методів комп'ютерної обробки надвеликих за обсягом даних в інформаційних середовищах різноманітного призначення.

Предмет навчальної дисципліни — сукупність підходів, інструментів і методів обробки структурованих і неструктурованих даних великих обсягів, а також методи машинного та глибокого навчання, що дозволяють видобувати з таких даних результати, придатні для прийняття рішень.

Результати навчання. В результаті вивчення дисципліни студенти повинні продемонструвати такі програмні результати навчання:

ПРН 1	Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп'ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп'ютерних наук та на межі галузей знань.
ПРН 2	Мати спеціалізовані уміння/навички розв'язання проблем комп'ютерних наук, необхідні для проведення досліджень та/або провадження інноваційної діяльності з метою розвитку нових знань та процедур.
ПРН 4	Управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів.
ПРН 6	Розробляти концептуальну модель інформаційної або комп'ютерної системи.
ПРН 7	Розробляти та застосовувати математичні методи для аналізу інформаційних моделей.
ПРН 8	Розробляти математичні моделі та методи аналізу даних (включно з великими).
ПРН 9	Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).
ПРН 11	Створювати нові алгоритми розв'язування задач у сфері комп'ютерних наук, оцінювати

	їх ефективність та обмеження на їх застосування.
ПРН 12	Проєктувати та супроводжувати бази даних та знань.
ПРН 14	Тестувати програмне забезпечення.
ПРН 15	Виявляти потреби потенційних замовників щодо автоматизації обробки інформації.

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

Пререквізити дисципліни.
Вивчення на попередньому рівні вищої освіти дисциплін з напрямів:

- бази даних,
- аналіз даних;
- зберігання даних.

Постреквізити дисципліни. Результати навчання використовуються під час виконання магістерської дисертації, а також у дисциплінах, пов'язаних з інтелектуальним аналізом даних, проєктуванням інформаційних систем та науково-дослідною роботою за темою дисертації.

3. Зміст навчальної дисципліни

Тема 1. Великі дані: поняття та інфраструктура.
Тема 2. Підготовка даних та класичне машинне навчання.
Тема 3. Нейронні мережі та глибоке навчання.
Тема 4. Сучасні застосування та машинне навчання на великих даних.

4. Навчальні матеріали та ресурси

1. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow. 3rd edition. O'Reilly Media, 2022, 861 p.
2. Prince S. J. D. Understanding Deep Learning. MIT Press, 2023, 541 p. Режим доступу: https://udlbook.github.io/udlbook/
3. Zhang A., Lipton Z., Li M., Smola A. Dive into Deep Learning. Cambridge University Press, 2023. Режим доступу: https://d2l.ai/
4. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016, 800 p.
5. Kleppmann M. Designing Data-Intensive Applications. O'Reilly Media, 2017, 616 p.
6. Chambers B., Zaharia M. Spark: The Definitive Guide. O'Reilly Media, 2018, 601 p.
7. Ланде Д. В. Оброблення надвеликих масивів даних (Big Data) [Електронний ресурс]: навчальний посібник для використання у навчальному процесі з підготовки фахівців другого (магістерського) рівня вищої освіти зі спеціальності 122 «Комп'ютерні науки» / Д. В. Ланде, І. Ю. Субач, А. Я. Гладун; КПІ ім. Ігоря Сікорського. Електронні текстові дані. Київ: КПІ ім. Ігоря Сікорського, 2021. 168 с.

Додаткова література

1. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006, 738 p.
2. Raschka S., Liu Y., Mirjalili V. Machine Learning with PyTorch and Scikit-Learn. Packt Publishing, 2022, 770 p.
3. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.
4. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. Proceedings of OSDI, 2004.

5. Armbrust M., Ghodsi A., Xin R., Zaharia M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. Proceedings of CIDR, 2021.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Тема 1. Великі дані: поняття та інфраструктура

Лекція 1. Основні аспекти та складові елементи трактування поняття «Великі дані». Характеристики великих даних: обсяг, швидкість, різноманітність, достовірність, цінність. Сфери застосування надвеликих масивів даних. Проблема масштабування. Базові компоненти конвеєра аналізу даних. Машинне навчання, його історичний розвиток і сучасний стан. Приклади прикладних задач, які використовують методи машинного навчання.

Самостійна робота студентів. Розглянути генезис поняття «Великі дані». Дібрати та описати прикладну задачу з обраної галузі, що потребує обробки надвеликих масивів даних.

Лекція 2. Розподілене зберігання даних. Архітектура HDFS: NameNode і DataNode, блоки, реплікація, конвеєр запису та читання даних. Забезпечення відмовостійкості. Об'єктні сховища як альтернатива розподіленій файловій системі. Менеджер ресурсів YARN.

Самостійна робота студентів. Ознайомитися зі структурою менеджера ресурсів YARN. Порівняти HDFS з об'єктними сховищами за вартістю, латентністю та моделлю узгодженості.

Лекція 3. Теорема CAP та її практичне трактування. Компроміс між узгодженістю та доступністю в умовах мережевого розділення. Моделі даних NoSQL: ключ-значення, документні, колоночні (wide-column), графові. Критерії вибору сховища під задачу.

Самостійна робота студентів. Ознайомитися з принципами NoSQL баз даних. Розглянути колоночну базу даних HBase та особливості її моделі даних.

Лекція 4. Модель обчислень MapReduce: етапи map, shuffle, reduce; типові шаблони проектування. Apache Spark: RDD і DataFrame, відкладені обчислення, граф виконання. Архітектури сховищ: Data Warehouse, Data Lake, Lakehouse. Схема при записі та схема при читанні. Колонковий формат Parquet. Транзакційний рівень Delta Lake.

Самостійна робота студентів. Ознайомитися з шаблонами проектування MapReduce. Опрацювати підходи до збереження та обробки надвеликих масивів даних у Spark.

Тема 2. Підготовка даних та класичне машинне навчання

Лекція 5. Постановка задач машинного навчання. Ознаки та цільова змінна. Підготовка даних: очищення, обробка пропусків, масштабування, кодування категоріальних ознак. Витік даних. Розбиття вибірок, крос-валідація. Аугментація. Метрики якості класифікації та регресії.

Самостійна робота студентів. Опрацювати вплив масштабування ознак на різні сімейства моделей. Розглянути стратифіковане розбиття вибірок для незбалансованих даних.

Лекція 6. Лінійна регресія з однією та з декількома змінними. Оціночна функція. Метод градієнтного спуску, вибір коефіцієнта навчання. Нормальні рівняння. Логістична регресія, границі рішень, регуляризація, багатокласова класифікація методом One-vs-all.

Самостійна робота студентів. Метод градієнтного спуску в логістичній регресії. Проблема недостатнього та надмірного навчання, способи її діагностики.

Лекція 7. Дерева рішень: критерії розбиття, ентропія та індекс Джині, обрізання дерева. Класифікація та регресія на основі дерев рішень. Ансамблеві методи: беггінг, випадковий ліс, градієнтний бустинг. Оцінювання важливості ознак.

Самостійна робота студентів. Порівняти випадковий ліс і градієнтний бустинг за стійкістю до перенавчання та вимогами до підготовки даних.

Лекція 8. Метод опорних векторів. Цілі оптимізації, максимізація зазору. Постановка та формальний опис задачі. Лінійно неподільна вибірка, м'який зазор. Застосування ядрових функцій.

Самостійна робота студентів. Ознайомитися з найбільш поширеними ядрами для методу опорних векторів та впливом їх гіперпараметрів на границю рішень.

Лекція 9. Навчання без учителя. Кластеризація методом к-середніх, оптимізація та вибір кількості кластерів, метрики відстаней. Метод головних компонент: діагоналізація коваріаційної матриці, сингулярний розклад матриці даних, відбір компонент. Метод колаборативної фільтрації та рекомендаційні системи.

Самостійна робота студентів. Властивості й обмеження методу головних компонент, відбір компонент за правилом Кайзера. Типи колаборативної фільтрації та обмеження на застосування методу.

Тема 3. Нейронні мережі та глибоке навчання

Лекція 10. Лінійна відокремлюваність. Персептрон, його навчання та обмеження. Багатошаровий персептрон. Функції активації та їх вплив на навчання глибоких мереж. Презентація даних у нейронних мережах.

Самостійна робота студентів. Порівняти функції активації за поведінкою похідних. Розглянути проблему затухання градієнта.

Лекція 11. Прямий і зворотний прохід. Метод зворотного поширення похибки. Функції втрат. Оптимізатори: SGD, momentum, Adam. Планування швидкості навчання. Регуляризація: dropout, weight decay, нормалізація. Параметри та гіперпараметри моделі.

Самостійна робота студентів. Ознайомитися зі стратегіями планування швидкості навчання та практикою підбору гіперпараметрів.

Лекція 12. Згорткові нейронні мережі. Просторові ієрархії ознак, операції згортки та підвибірки, рецептивне поле. Типові архітектури згорткових мереж. Задачі детекції об'єктів, придушення немаксимумів.

Самостійна робота студентів. Розглянути згортку 1×1 та її роль у зменшенні обчислювальної складності. Ознайомитися з метриками якості детекції.

Лекція 13. Моделювання послідовностей. Рекурентні нейронні мережі, архітектури LSTM та GRU. Навчання на часових рядах: нормалізація, ковзне вікно, teacher forcing. Обмеження рекурентних архітектур.

Самостійна робота студентів. Особливості підготовки часових рядів для уникнення витоку даних. Багатокрокове прогнозування.

Лекція 14. Архітектура трансформера. Механізм уваги, багатоголова увага, позиційне кодування. Кодувальник і декодувальник. Авторегресивна генерація, маскування. Передумови масштабованості архітектури.

Самостійна робота студентів. Ознайомитися з обчислювальною складністю механізму уваги та підходами до її зниження.

Тема 4. Сучасні застосування та машинне навчання на великих даних

Лекція 15. Мультимодальні моделі та токенізація. Підходи до генеративного моделювання, керування генерацією. Машинне навчання на великих даних: розподілене навчання, паралелізм за даними та за моделлю, бібліотека Spark MLlib. Батчевий та онлайн inference. Дрейф даних і моніторинг моделей у промисловій експлуатації.

Самостійна робота студентів. Розглянути підходи до розподіленого навчання. Вирішення задач машинного навчання з використанням бібліотеки MLlib.

Практичні заняття

Практичне заняття 1. Постановка задачі аналізу даних: формулювання гіпотези, вибір цільової метрики, планування експерименту та схеми валідації.

Практичне заняття 2. Вибір архітектури зберігання даних під задачу: порівняння реляційних, NoSQL та озерних рішень на конкретних сценаріях.

Практичне заняття 3. Аналіз помилок моделей: діагностика недо- та перенавчання, інтерпретація матриці помилок і метрик якості.

Практичне заняття 4. Розбір сучасних архітектур глибокого навчання за першоджерелами: обговорення обмежень і області застосовності.

6. Самостійна робота здобувача вищої освіти

Самостійна робота здобувача вищої освіти (68 годин) передбачає підготовку до лекційних, практичних та лабораторних занять у середовищі Google Colab, опрацювання рекомендованої літератури, виконання завдань за темами дисципліни та підготовку до контрольних заходів. Розподіл годин: підготовка до лекцій – 15 годин; підготовка до практичних занять – 4 години; підготовка та виконання лабораторних робіт – 21 година (по 3 години на роботу); підготовка до модульної контрольної роботи – 6 годин; підготовка до екзамену – 22 години. Разом – 68 годин.

Тема 1. Великі дані: поняття та інфраструктура

Розглянути генезис поняття «Великі дані» та прикладні задачі, що потребують обробки надвеликих масивів даних. Опрацювати особливості екосистеми Hadoop і структуру менеджера ресурсів YARN. Ознайомитися з шаблонами проектування MapReduce. Виконати огляд технологій зберігання великих даних, принципів NoSQL баз даних та колоночної бази даних HBase. Опрацювати підходи до збереження й обробки даних у Spark, зокрема колонковий формат Parquet і транзакційний рівень Delta Lake.

Тема 2. Підготовка даних та класичне машинне навчання

Опрацювати підходи до підготовки даних: обробку пропусків, масштабування, кодування ознак, запобігання витоку даних. Детальніше ознайомитися з поняттями класифікації та регресії, методом градієнтного спуску та регуляризацією. Розглянути ансамблеві методи та порівняти їх з одиночними деревами рішень. Ознайомитися з найбільш поширеними ядрами методу опорних векторів. Опрацювати метрики відстаней у задачах кластеризації, правило Кайзера для відбору головних компонент та типи колаборативної фільтрації.

Тема 3. Нейронні мережі та глибоке навчання

Опрацювати властивості функцій активації та проблему затухання градієнта. Детальніше ознайомитися з методом зворотного поширення похибки, оптимізаторами та стратегіями планування швидкості навчання. Розглянути типові архітектури згорткових мереж і метрики якості детекції. Опрацювати особливості підготовки часових рядів та багатокрокове прогнозування. Ознайомитися з обчислювальною складністю механізму уваги.

Тема 4. Сучасні застосування та машинне навчання на великих даних

Ознайомитися з мультимодальними моделями та підходами до токенізації. Опрацювати методи розподіленого навчання, паралелізм за даними та за моделлю. Вирішити навчальну задачу машинного навчання з використанням бібліотеки Spark MLlib. Розглянути причини дрейфу даних та підходи до моніторингу моделей у промисловій експлуатації.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Відвідування лекційних та лабораторних занять є обов'язковим за винятком поважних причин (хвороби, форс-мажорних обставин).

В разі пропущення занять з поважних причин викладач надає можливість студенту виконати усі або деякі лабораторні завдання (винятком є виконання деяких завдань у зв'язку із закінченням навчального процесу).

В разі пропущення занять без поважних причин, а також через порушення граничного терміну виконання завдання (deadline) студент може отримати не більше 80% від максимальної оцінки за відповідне завдання.

Протягом семестру студенти:

- виконують та захищають лабораторні роботи у відповідні терміни (на кожну лабораторну роботу відводиться два тижні для здачі),
- пишуть модульну контрольну роботу,
- повинні позитивно закрити два календарні контролі (тиждень 7–8 та тиждень 14–15),
- по закінченні навчального процесу складають екзамен.

Політика та принципи академічної доброчесності визначені у розділі 3 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Норми етичної поведінки студентів і працівників визначені у розділі 2 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Використання систем генеративного штучного інтелекту під час виконання лабораторних робіт дозволяється за умови, що студент декларує факт їх застосування у звіті, здатен пояснити призначення кожного фрагмента коду та обґрунтувати обраний метод. Нездатність пояснити власний код або інтерпретувати отримані результати розглядається як порушення академічної доброчесності незалежно від працездатності програми.

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

Система рейтингових (вагових) балів та критерії оцінювання

1) Робота на лекціях

На лекціях може бути проведено бліцопитування студентів. Такі опитування проводяться на довільних лекціях 5 разів протягом семестру, наприкінці лекції. Ваговий бал за вірну відповідь 1. Максимальна кількість балів, що може отримати кожен студент за семестр - 5.

2) Лабораторні роботи

Максимальна кількість балів за усі виконані лабораторні роботи дорівнює 50 балів. Розподіл балів серед лабораторних робіт наступний:

№ з/п	Назва лабораторної роботи	Кількість балів
1	Аналіз галузевого кейсу за характеристиками «великих даних» та профілювання набору даних	6
2	Вирішення задачі регресії	7
3	Порівняльний аналіз моделей класифікації: логістична регресія, метод опорних векторів, випадковий ліс	8
4	Вирішення задачі кластеризації та реалізація методу головних компонент	7

5	Реалізація рекомендаційної системи з використанням методу колаборативної фільтрації	7
6	Класифікація зображень за допомогою згорткової нейронної мережі	8
7	Обробка даних, що не вміщуються в оперативну пам'ять: базові операції MapReduce та PySpark	7
Разом		50

Критерії оцінювання:

Виконання лабораторної роботи:

- виконана своєчасно (протягом двох тижнів з моменту видачі), у повному обсязі відповідний бал згідно номеру лабораторної роботи;
- виконана із запізненням — знімається від 10 до 30% від максимальної кількості балів залежно від терміну запізнення;
- виконана не самостійно, із запізненням знімається 50% від максимальної кількості балів;
- невиконана протягом відведеного часу 0 балів.

3) Модульна контрольна робота

Ваговий бал – 10. Максимальна кількість балів за контрольну роботу дорівнює 10 балів.

Якість виконання роботи:

- всі відповіді вірні та повні -10 балів,
- у відповідях допущені несуттєві неточності - 8 балів,
- половина відповідей вірна - 5 балів,
- відповіді з суттєвими неточностями, але без критичних помилок – 2 бали.

4) Складання екзамену

Максимальний ваговий бал rекз = 35

Умови позитивного календарного контролю.

Для отримання „зараховано” з першого календарного контролю (тиждень 7–8) студент має набрати не менше ніж 14 балів (за умови, що на цей момент згідно з календарним планом контрольних заходів „ідеальний” студент має отримати 6+7+8+3 = 24 бали).

Для отримання „зараховано” з другого календарного контролю (тиждень 14–15) студент має набрати не менше ніж 35 балів (за умови, що на цей момент згідно з календарним планом контрольних заходів „ідеальний” студент має отримати 24+7+7+8+10 = 56 балів).

Умови допуску до екзамену.

Необхідною умовою допуску до екзамену є зарахування усіх лабораторних робіт та виконання модульної контрольної роботи, а також стартовий рейтинг (R_C) не менше 40 балів. Для отримання екзамену з кредитного модуля "автоматом" потрібно мати рейтинг не менш ніж 60 балів, а також зараховане виконання всіх завдань лабораторних робіт.

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру (шкала рейтингу) складає:

R = rлек + rлаб + rмод + rекз = 5 + 50 + 10 + 35 = 100 балів.

Стартовий рейтинг становить RC = rлек + rлаб + rмод = 65 балів.

Рейтинг екзамену дорівнює 35 балів.

Таким чином, рейтингова шкала з кредитного модуля складає

R=65+35=100 балів.

Для отримання студентом відповідних оцінок (ECTS та традиційних) його рейтингова оцінка RD переводиться згідно таблиці:

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
R≤ 59	Незадовільно
RC < 40 або не виконані інші умови допуску	Не допущено

Робочу програму навчальної дисципліни (силабус):

Складено асистентом, доктором філософії, Здором Костянтином Андрійовичем

Ухвалено кафедрою ЦТЕ (протокол № 22 від 17.06.2026)

Погоджено науково-методичною комісією НН ІАТЕ (протокол № 9 від 26.06.2026 р.)