МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
Кафедра системного проектування

Затверджено
На засіданні кафедри системного проектування
факультету електроніки та комп'ютерних технологій
Львівського національного університету імені Івана Франка
(протокол № 1 від 28.08 2023 р.)

Завідувач кафедри:

_____ Роман ШУВАР

Силабус з навчальної дисципліни "Основи інженерії даних", що викладається в межах ОПП "Комп'ютерні науки" першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 122 – Комп'ютерні науки

Львів 2023 р.

Назва дисципліни	Основи інженерії даних
Адреса викладання дисципліни	Корпус факультету електроніки та комп'ютерних технологій, Львівський національний університет імені Івана Франка, вул. Драгоманова 50, м. Львів, 79005, вул. Ген. Тарнавського 107, м. Львів, 79011
Факультет та кафедра, за якою закріплена дисципліна	Факультет електроніки та комп'ютерних технологій, кафедра системного проектування
Галузь знань, шифр та назва спеціальності	12 – інформаційні технології 122 – Комп'ютерні науки
Викладачі дисципліни	Демків Л.С., канд. фіз.-мат. наук, доцент
Контактна інформація викладачів	lidiya.demkiv@lnu.edu.ua, https://electronics.lnu.edu.ua/employee/demkiv-l-s
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекційних занять (за попередньою домовленістю). Для погодження часу онлайн консультацій слід писати на електронну пошту викладача.
Сторінка дисципліни	https://moodle.elct.lnu.edu.ua/course/view.php?id=79
Інформація про дисципліну	Дисципліна «Основи інженерії даних» є вибірковою дисципліною з спеціальності 122 – Комп'ютерні науки програми " Комп'ютерні науки", яка викладається в 5-му семестрі в обсязі 3,5 кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Мета та цілі дисципліни	Мета: реалізувати знайомство студента із способами створення цілісних та чистих даних, теоретичними принципами галузі інженерії даних та практичними технологіями, які найбільш затребувані в інженерії даних, а також способами розробки систем для ефективного збору та зберігання великих обсягів даних з різних джерел. Цілі: забезпечити знайомство студентів з особливостями обробки та зберігання різних типів даних, варіантами побудови сховищ даних; вивчити алгоритми попередньої обробки даних; опанувати теоретичний матеріал і практичне оволодіння сучасними графічно-інформаційними технологіями, комп'ютерними та програмними засобами створення цілісних даних, подання їх в графічній формі; вивчити методи і алгоритми обробки даних, визначити статистичні параметри даних; ознайомити з базовими концепціями обробки даних, які дозволять правильно структурувати дані для подальшого їхнього опрацювання, візуалізації і моделювання, управління програмною інфраструктурою та інтерфейсом систем обробки даних, теорією і проектуванням систем обробки даних. Розробка інфраструктури для інтеграції та підтримки інструментів аналітики даних, таких як системи Business Intelligence (BI) та інші.
Література для вивчення дисципліни	Основна література: 1. Технології оброблення великих даних: конспект лекцій з дисципліни «Технології оброблення великих даних» [Електронний ресурс] : навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» (освітня програма «Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем»)/ Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 5,55 Мбайт). – Київ: КПІ ім. Ігоря Сікорського, 2019. – 227 с

	2. Han, Jiawei. Data mining : concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed. ISBN 978-0-12-381479-1 Chapter 3. Data preprocessing Michael R. Brzustowicz Data Science with Java Practical Method for scientists and engineers /Michael R. Brzustowicz. – O'REILLY, 2017. – 311p. 3. Методичні вказівки до виконання лабораторних робіт з дисципліни "Інженерія прикладних інтелектуально-орієнтованих програмних продуктів" для студентів спеціальностей 121 "Інженерія програмного забезпечення" та 122 "Комп'ютерні науки та інформаційні технології" (всіх форм навчання) / В.М. Льовкін. – Запоріжжя : ЗНТУ, 2016. – 80 с. 4. Edward L.Robinson Data Analysis for Scientists and Engineers // Pricenton University Press, 2016, - P.408, ISBN 9781400883066 5. Sayan M., Pratip S. Advanced Data Analytics Using Python: With Architectural Patterns, Text and Image Classification, and Optimization Techniques// APress, 2022. – 249p. 6. Fabio Nelli Python Data Analytics: With Pandas, NumPy, and Matplotlib 3rd ed. Edition // Apress, 2023. – 466p. 7. Glaucia Esppenchutz Data Ingestion with Python Cookbook: A practical guide to ingesting, monitoring, and identifying errors in the data ingestion process // Packt Publishing, 2023. – 414. 8. A.J.Heyley, D.Wolf Learn Data Analysis with Python: Lessons in Coding // Apress, 2018. – 97p. Додаткова література 1. Extract, transform, and load (ETL) // Електронний ресурс. Режим доступу: https://docs.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl 2. Data Lake vs. Data Warehouse: What's the Difference? // Електронний ресурс. Режим доступу: https://www.coursera.org/articles/data-lake-vs-data-warehouse 3. Databases vs. Data Warehouses vs. Data Lakes // Електронний ресурс. Режим доступу: https://www.mongodb.com/databases/data-lake-vs-data-warehouse-vs-database 4. Data Engineering: матеріали для самопідготовки// Електронний ресурс. Режим доступу: https://training.epam.ua/ua/blog/131
Обсяг курсу	64 години аудиторних занять. З них 32 години лекцій, 32 години лабораторних робіт та 41 годин самостійної роботи
Очікувані результати навчання	Після завершення цього курсу студент буде використовувати сучасні програмні засоби для попередньої обробки неідеальних реальних даних, запису даних у відповідні структури та сховища даних, моделювання та інтеграції даних; реалізовувати інтерактивні візуалізації даних; проводити необхідну попередню обробку даних для отримання чистих даних; визначати тип задачі аналізу та вирішувати її адекватно обраним методом з оптимально визначеними параметрами; оцінювати результати; робити змістовні висновки та інтерпретацію опрацювання даних.
Ключові слова	Структури даних, ETL, DATA WAREHOUSE, DATA LAKE
Формат курсу	Очний

	Проведення лекцій, лабораторних робіт та консультації для кращого розуміння тем
Теми	Див. СХЕМА КУРСУ
Підсумковий контроль, форма	Залік в кінці семестру
Пререквізити	Для вивчення курсу студенти потребують базових знань з дисциплін «Вища математика», «Дискретна математика», «Теорія обчислень, алгоритми та структури даних», «Бази даних».
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Презентація, лекції, лабораторні роботи, обговорення, дискусія.
Необхідне обладнання	Для проведення лекційних занять: комп'ютер з мінімальними характеристиками: процесор Intel Core i3 або аналогічний фірми AMD, з 4ГБ або більше оперативної пам'яті, доступ до мережі Internet, засоби мультимедіа: мультимедійний проектор та екран. Для проведення лабораторних занять необхідно: 1. комп'ютерна лабораторія з 12-14 робочими місцями, комп'ютери з мінімальними характеристиками: процесор Intel Core i3 або аналогічний фірми AMD, з 4ГБ або більше оперативної пам'яті, доступ до мережі Internet. 2. програмне забезпечення включає в себе ОС Windows 10 або дистрибутив Linux, середовище Moodle, Python та середовище для програмування в ньому PyCharm, Jupyter Notebook або аналогічні.
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться упродовж семестру за 100-бальною шкалою. Бали нараховуються за такими видами робіт з наступним співвідношенням: • лабораторні роботи: 50% семестрової оцінки; максимальна кількість балів 50. • Виконання теоретичних, індивідуальних і практичних завдань: 50% семестрової оцінки; максимальна кількість балів 50. Загалом упродовж семестру 100 балів. Оцінювання лабораторних робіт (10 лабораторних робіт, максимальна кількість балів: 50) відбувається шляхом оцінки роботи студента під час проведення лабораторної роботи в аудиторії (0-2 балів за одну роботу) та захисту звіту по виконаній лабораторній роботі (1-3 бали за одну роботу). Бали оцінювання лабораторних робіт нараховуються за наступним співвідношенням: на парі (0-2 бали) за виконання фрагменту завдання до лабораторної роботи 0 балів – студент не виконав завдання 1 бал – студент частково виконав завдання 2 бали – студент повністю виконав завдання на парі за захист звіту (0-3 бали) 3 – студент в повному обсязі володіє навчальним матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 2 – студент не досить добре розуміє розглянутий матеріал та написаний

	ним код програми, вагається та надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує неточно, або з помірними недоліками; 1 - студент погано розуміє розглянутий матеріал та написаний ним код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми функціонує з суттєвими недоліками; 0 - студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. Нарахування балів - 20 балів за проходження тесту - по 15 балів за виконання 2 практичних індивідуальних завдань. Бали оцінювання кожного екзаменаційного теоретичного питання нараховуються за наступним співвідношенням: 14-15 - розглянуте питання викладено в повному обсязі, правильно, обґрунтовано, логічно, містить аналіз і систематизацію, аргументовані висновки. Засвідчено глибоке володіння матеріалом. Наведено приклади використання теоретичного матеріалу. Можуть бути присутні несуттєві описки та невідповідності; 10-13 – у відповіді висвітлено значна частина питання. Виявлено знання і розуміння основних положень навчальної дисципліни, проте присутні неточності та/або невідповідності основній темі. Приклади використання теоретичного матеріалу відсутні; Код є високоефективним та оптимізованим, без помилок, зроблені висновки, які логічно випливають з отриманих результатів 7-9 – у відповіді відстежується загальне розуміння розглянутої теми. Виявлені множинні неточності та невідповідності, пояснення наведених формул відсутні чи частково помилкові; Код ефективний, але може бути дещо оптимізованим. Загалом зроблені висновки правильно пояснюють отримані результати. 4-7 – з відповіді студента зрозуміло, що він погано розуміє розглянуту тему. Виявлені суттєві неточності та невідповідності. Наведені факти майже не відповідають темі; Код має помітні помилки, але деякі частини працюють правильно. Є висновки, але вони можуть бути більш детальними або зв'язаними з кодом. 0 – 3 – студент взагалі не розуміє розглянуту тему. Тему не розкрито, кількість викладеного матеріалу не відповідає загальним нормам. : Код неефективний, потребує значної оптимізації. Код містить серйозні помилки, які суттєво впливають на його функціональність. **Критерії оцінювання результатів неформальної освіти:** Нарахування балів відбувається за написання студентом тез доповідей на конференціях, наукових статей, участь у діяльності наукових гуртків, участь у наукових семінарах та круглих столах, конкурсах, участь у заходах неформальної освіти за отримання сертифікатів про проходження навчання на різних освітніх платформах (Coursera, Prometheus тощо), курсах на провідних IT компаніях за тематикою навчальної дисципліни. Кількість балів визначається відсотком покриття результатів відповідної активності до вимог результатів навчання з навчальної дисципліни. _____. **Контрольні заміри проводяться у формі тестових завдань. Академічна доброчесність**: Очікується, що лабораторні та контрольні

	роботи студентів будуть їх оригінальними дослідженнями чи міркуваннями. Відсутність посилань на використані джерела, фабрикування джерел, списування, втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в роботі студента є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману. **Відвідування занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції і лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати заняття. Студенти зобов'язані дотримуватися усіх термінів, визначених для виконання усіх видів робіт, передбачених курсом. **Література.** Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих. **Політика виставлення балів.** Враховуються бали, набрані на поточному тестуванні, самостійній роботі та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час лабораторного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях, не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін. Жодні форми порушення академічної доброчесності не толеруються.
Питання до модульного контролю	Перелік питань та завдань для проведення підсумкової оцінки знань 1. Індустрія опрацювання зростаючої кількості даних у світі. 2. Професії, які працюють з даними. Особливості завдань, які вони виконують. 3. Статистичні параметри даних. 4. Нормальний розподіл даних. 5. Види графіків для візуалізації даних. 6. Поняття кореляції даних. Матриця кореляції. 7. Поняття дослідницького аналізу даних. 8. Поняття класифікації даних. 9. Кроки для реалізації моделі класифікації даних. 10. Алгоритм класифікації даних DT. 11. Алгоритм RF. 12. Алгоритм PCA. 13. Кластеризація даних. Поняття відстані між об'єктами кластеризації. 14. Методи кластеризації даних. 15. Алгоритм k-means для кластеризації даних. 16. Поняття регресійного аналізу. 17. Рекомендаційні системи та методи їх реалізації. 18. Види графів. Види графіків для зображення графів. 19. Json формат даних. 20. Поняття ETL процесів. 21. Поняття сховища даних DWH. Практичні завдання • Провести дослідницький аналіз даних • Провести кластерний аналіз даних • Провести класифікацію даних • Провести BI аналіз даних

Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

СХЕМА КУРСУ

Тиж.	Тема, план, короткі тези	Форма діяльності (заняття)	Література. Ресурси в Інтернеті	Завдання (лабораторна робота), год	Термін виконання
1	Поняття даних. Визначення кількості даних та проблеми швидкого збільшення кількості даних в сучасному світі. Професії, що працюють з даними. Відкриті дані — це цінний ресурс, який допомагає посилити цифрову та "реальну" економіки країни. Засоби вилучення, перетворення і завантаження даних. Поняття ETL, pipeline, datalake, data warehouse.	Лекція	1, 2, 5	Вступне заняття. Інструкція з техніки безпеки. Методи бібліотеки pandas для роботи з даними	2 тиж. семестру
2	Формати зберігання даних. Джерела даних: таблиці,файли, бази даних, web-сервіси. Зчитування та запис файлових даних. Типи даних за шкалами вимірювання.	Лекція	1, 3, 5, 6	Вивчення методів бібліотеки pandas (python) для роботи з даними: сортування, групування, фільтрація, об’єднання даних. Основи BI аналітики.	6 тиж. семестру
3	Візуалізація даних як етап аналізу даних	Лекція	1, 5	Підготовка даних до візуалізації. Візуалізація даних.	6 тиж. семестру
4	Коваріація та кореляція даних. Кореляційний аналіз кількісних ознак. Побудова рекомендаційних систем.	Лекція	2, 3	Коваріація та кореляція даних.	7 тиж. семестру
5	Статистичний аналіз даних. Обчислення описової статистики даних. Статистичні моменти даних. Роботи з великою кількістю даних (обновлення статистики даних) Нормування та стандартизація даних	Лекція	1, 2, 5, 6	Статистичний аналіз даних	7 тиж. семестру
6	Розподіли числових даних. Нормальний розподіл даних. Види розподілів. Перевірка гіпотез. Види гіпотез. P-		2, 6	Захисне заняття	7 тиж. семестру

	value. Параметричні та непараметричні тести.				
7	Дослідницький аналіз даних. Етапи дослідницького аналізу даних.	Лекція	2, 5,	Дослідницький аналіз даних.	9 тиж. семестру
8	Створення даних для бізнес аналітики (BI)	Лекція	7	Побудова дашбордів за допомогою Tabloau Підсумкове заняття ЗМ1	10 тиж. семестру
9	Підготовка даних до класифікації. Редукція та класифікація даних. Метод головних компонент.	Лекція	2,4,7	Підготовка даних до класифікації та ML	11 тиж. семестру
10	Побудова дерев класифікації. Random Forest. Порівняння методів класифікації.	Лекція	1, 2, 4	Редукція та класифікація даних	12 тиж. семестру
11	Доступ до баз даних за допомогою модулів python. API Python для роботи з базами даних MySQL, SQLite, Принципи побудови DataLake.	Лекція	1, 2, 4	Доступ до баз даних за допомогою модулів python	14 тиж. семестру
12	Поняття ETL. Принципи побудови сховищ даних DataWareHouse.		6,7	Моделювання побудови сховищ даних DataWareHouse	14 тиж. семестру
13	Робота з нереляційними базами даних. JSON-формат та особливості його використання. Створення та парсинг формату. Типи даних json та масиви. Особливості зчитування json у python. Підготовка даних до кластеризації.	Лекція	2, 5,	JSON-формат та особливості його використання..	15 тиж. семестру
14	Графові бази даних. Графи та структури даних на графах. Представлення та візуалізація даних за допомогою графів	Лекція	2, 3, 4, 5	Графи та структури даних на графах.	16 тиж. семестру
15	Особливості роботи з сирими raw даними. Чисельні системи представлення даних та частота дискретизації даних	Лекція	2. 4. 5	Особливості роботи з сирими raw даними.	16 тиж. семестру
16	Pipeline - трубопровід для опрацювання даних. Схеми побудови трубопроводів. Зберігання та оброблення даних в розподілених файлових системах. Особливості роботи з BigData.	Лекція	4, 5, 6, 7	Підсумкове заняття ЗМ2	16 тиж. семестру