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 30. 08. 2022 р.)

Завідувач кафедри:
_____ Роман ШУВАР

Силабус з навчальної дисципліни "Методи та технології опрацювання даних", що викладається в межах ОПП " Високопродуктивний комп'ютинг " першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 121 – Інженерія програмного забезпечення

Львів 2022 р.

Назва дисципліни	Методи та технології опрацювання даних
Адреса викладання дисципліни	м. Львів, вул. Драгоманова, 50
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 121 Інженерія програмного забезпечення (ВПК)
Викладачі дисципліни	Юзевич Володимир Миколайович, професор Ляшкевич Василь Яремович, доцент
Контактна інформація	volodymyr.yuzevych@lnu.edu.ua vasyl.lyashkevych@lnu.edu.ua
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекційних занять (за попередньою домовленістю). Також можливі он-лайн консультації через MS Teams або систему електронного навчання Moodle. Для погодження часу онлайн консультацій слід писати на електронну пошту викладача.
Сторінка дисципліни	https://moodle.elct.lnu.edu.ua/course/view.php?id=243
Інформація про дисципліну	Дисципліна «Методи та технології опрацювання даних» є нормативною дисципліною з циклу професійної та практичної підготовки за блоками вибіркових дисциплін з спеціальності 121 Інженерія програмного забезпечення для освітньої програми «Високопродуктивний комп'ютинг», яка викладається в 5 семестрі в обсязі 3,5 кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено таким чином, щоб надати учасникам необхідні знання, щоб оволодіти базовими поняттями, пов'язаними з алгоритмами, методами та засобами опрацювання даних, побудові конвеєрів даних, використання метрик та засобів оцінки даних, архітектури даних та інтерпретація стурктурованих і неструктурованих даних. Саме тому у дисципліні подано огляд базових понять та інструментів для опрацювання даних, так і засобів, які потрібні для вирішення типових завдань при побудові конвеєрів даних, аналізу та візуалізації даних.
Мета та цілі дисципліни	Метою вивчення нормативної дисципліни "Методи та технології опрацювання даних" є оволодіння базовими поняттями, теоретичними знаннями та практичними навичками опрацювання даних, візуалізації даних, побудови конвеєрів для аналізу і перетворення даних з подальшим використанням в різних галузях людської діяльності з метою вирішення різного роду задач і бізнес проблем.

Література для вивчення дисципліни	Основна література: 1. Paul Crickard. Data Engineering with Python - Birmingham: Packt Publishing, 2020. - 337 p. - ISBN 978-1-83921-418-9. 2. Wes McKinney. Python for Data Analysis - Sebastopol: O'Reilly Media, 2018. - 522 p. - ISBN 978-1-491-95766-0. 3. Joakim Sundnes. Introduction to Scientific Programming with Python - Lysaker: Simula SpringerBriefs, 2020, Volume 6. - ISBN: 978-3-030-50355-0. (eBook) 4. Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser. Data Structures & Algorithms in Python. Wiley: Courier Westford, 2013. - 748 p. (eBook) 5. Numpy community. Numpy User Guide. Release 1.18.4: May 24, 2020. - 166 p. 6. Dr. Ossama Embarak. Data Analysis and Visualization Using Python - Abu Dhabi: Apress Media LLC, 2018. - 374 p. - ISBN-13 (pbk): 978-1-4842-4108-0. 7. Massimo di Pierro. Annotated Algorithms in Python - Chicago: Experts4Solutions, 2017. - 227 p. - ISBN: 978-0-9911604-0-2. 8. Allen B. Downey. Think Stats. Exploratory Data Analysis in Python - Needham: Green Tea Press, 2014. - 244 p. 9. Jake VanderPlas. Python Data Science Handbook - Sebastopol: O`Reilly Media, 2017. - 517 p. - ISBN: 978-1-491-91205-8. 10. The Ultimate Guide to Basic Data Clearning: Atlan, 2014. - 66 p. 11. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining : concepts and techniques - Waltham: Elsevier, 2012. - 703 p. 12. Peter Bruce, Andrew Bruce, Peter Gedeck. Practical Statistics for Data Scientists. - Sebastopol: O`Reilly, 2020. - 329 p. - ISBN: 978-1-492-07294-2. 13. Brian Godsey. Think Like a Data Scientist. - Shelter Island: Manning Publications, 2017. - 299 p. - ISBN: 9781633430273. 14. Meher Krishna Patel. Pandas Guide. - May, 2020. - 62 p. 15. Aurelien Geron. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. - Sebastopol: O`Reilly, 2019. - 482 p. - ISBN: 978-1-492-03264-9. 16. Lewandowska, A.; Joachimiak-Lechman, K.; Kurczewski, P. A Dataset Quality Assessment—An Insight and Discussion on Selected Elements of Environmental Footprints Methodology. *Energies* 2021, *14*, 5004. https://doi.org/10.3390/en14165004 17. Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data Quality Assessment / Communications of the ACM, Volume 45, Issue 4, April 2002 pp. 211–218. - https://doi.org/10.1145/505248.506010 18. J. Bicevskis, Z. Bicevska, A. Nikiforova and I. Oditis, "An Approach to Data Quality Evaluation," *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 196-201, doi: 10.1109/SNAMS.2018.8554915. 19. Mats Bergdahl, Manfred Ehling, Eva Elvers and others. Handbook on Data Quality Assessment Methods and Tools. - Wiesbaden, 2007. - 139 p. 20. Mark Richards. Software Architecture Patterns . - Sebastopol: O`Reilly Media, 2015. - 45 p. - ISBN: 978-1-491-92424-2. 21. Dimensionality reduction [Режим доступу]: http://bioconductor.org/books/3.15/OSCA.basic/dimensionality-reduction.html 22. Data exploration with alluvial plots [Режим доступу]: https://www.datisticsblog.com/2018/10/intro_easyalluvial/#features 23. Khaled El Emam, Lucy Mosquera, Richard Hoptroff. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data: O`Reilly, 2020 24. Amazon. Lambda Architecture for Batch and Stream Processing. - AWS, 2018. - 12 p. 25. Tomcy John, Pankaj Misra. Data Lake for Enterprises. - Packt Publishing, 2017. - 855p.
Обсяг курсу	Кількість кретитів ЄКТС: 3.5 (105 год), з них: 64 годин аудиторних занять (лекції: 32 год, лабораторні: 32 год.) та 41 год. самостійної роботи.
Очікувані результати навчання	Після вивчення даного курсу здобувачі набудуть таких Загальних(ЗК)/Фахових(ФК) компетентностей та Програмних результатів навчання (ПРН): ЗК01. Здатність до абстрактного мислення, аналізу та синтезу. ЗК02. Здатність застосовувати знання у практичних ситуаціях. ЗК06. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

	ФК15. Здатність розробляти архітектури, модулі та компоненти програмних систем. ФК16. Здатність формулювати та забезпечувати вимоги щодо якості програмного забезпечення у відповідності з вимогами замовника, технічним завданням та стандартами. ФК18. Здатність аналізувати, вибирати і застосовувати методи і засоби для забезпечення інформаційної безпеки (в тому числі кібербезпеки). ФК19. Володіння знаннями про інформаційні моделі даних, здатність створювати програмне забезпечення для зберігання, видобування та опрацювання даних. ФК25. Здатність обґрунтовано обирати та освоювати інструментарій з розробки та супроводження програмного забезпечення. ФК26. Здатність до алгоритмічного та логічного мислення. ФК27. Здатність розробляти високопродуктивні програмні комплекси для вирішення завдань наук про дані, систем штучного інтелекту, вбудованих та інших інноваційних систем. ПРН01. Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідникові ресурси і знання з урахуванням сучасних досягнень науки і техніки. ПРН04. Знати і застосовувати професійні стандарти і інші нормативно-правові документи в галузі інженерії програмного забезпечення. ПРН05. Знати і застосовувати відповідні математичні поняття, методи доменного, системного і об’єктно-орієнтованого аналізу та математичного моделювання для розробки програмного забезпечення. ПРН09. Знати та вміти використовувати методи та засоби збору, формулювання та аналізу вимог до програмного забезпечення. ПРН10. Проводити передпроектне обстеження предметної області, системний аналіз об’єкта проектування. ПРН11. Вибирати вихідні дані для проектування, куруючись формальними методами опису вимог та моделювання. ПРН13. Знати і застосовувати методи розробки алгоритмів, конструювання програмного забезпечення та структур даних і знань. ПРН14. Застосовувати на практиці інструментальні програмні засоби доменного аналізу, проектування, тестування, візуалізації, вимірювань та документування програмного забезпечення. ПРН18. Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних. ПРН23. Вміти документувати та презентувати результати розробки програмного забезпечення. ПРН25. Вміти застосовувати сучасні технологічні рішення щодо розробки програмно-апаратних систем та їх компонентів. ПРН26. Знати засоби інтеграції, розгортання та підтримки спеціалізованих програмних компонентів, розроблених на

	основі інноваційних технологій для вирішення завдань високопродуктивних обчислень.
Ключові слова	Інженерія даних, опрацювання даних, інженерія ознак, конвеєри даних, візуалізація даних, Data Engineering, Data processing pipeline, feature engineering, feature importance, data visualization, a big data, Apache Spark, PySpark, Python.
Формат курсу	Проведення лекцій, лабораторних робіт та консультації для кращого розуміння тем проводиться у формі проектно-орієнтованого підходу з елементами дуальної освіти в компанії ГлобалЛоджик.
Теми	Див. СХЕМА КУРСУ
Підсумковий контроль, форма	Залік в кінці семестру
Пререквізити	Для вивчення курсу студенти потребують базових знань з дисциплін «Вища математика», «Прикладна статистика та ймовірнісні процеси», «Бази даних».
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Презентація, лекції, лабораторні роботи, обговорення, дискусія.
Необхідне обладнання	Мультимедійне обладнання, комп’ютерний клас, програми та сервіси MS Teams, Moodle, Python, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn та ін.

Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться упродовж семестру за 100-бальною шкалою, де враховано бали за два контрольні заміри по 35 балів за кожний модуль та 30 балів за складання заліку. Бали нараховуються за такими видами робіт з наступним співвідношенням: • контрольні заміри (2 модулі): 70% семестрової оцінки; максимальна кількість балів 70, а саме: - лабораторні роботи: 60% оцінки контрольного заміру; максимальна кількість балів 42 (15 лабораторних робіт). - теоретичний матеріал: 40% оцінки контрольного заміру; максимальна кількість балів 28 (2 тести по 14 балів кожний). • залік: 30% семестрової оцінки, максимально 30 балів. Оцінки за лабораторні заняття розподіляються наступним чином: виконання лабораторних завдань – 60 %, відповіді на запитання викладача – 40 %. Оцінки за **лабораторні роботи** розподіляються: **3** (0-2 бали за виконання, 0-1 бал за тестування/опитування) – 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** (0-1 бали за виконання, 0-1 бал за тестування/опитування) – 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** (0-1 бали за виконання, 0-1 балів за тестування/опитування) - студент погано розуміє розглянутий матеріал та написаний ним код програми, код програми не функціонує належним чином; **0** (0 балів за виконання, 0 балів за тестування/опитування) - студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. **Оцінювання залікових питань:** **10 балів** - розглянута тема відтворюється в повному обсязі, правильно, обґрунтовано, логічно, які містять аналіз і систематизацію, аргументовані висновки. Засвідчено глибоке володіння матеріалом. Наведені приклади коду повністю робочі та відповідають темі. Можуть бути присутні несуттєві помилки та невідповідності; **8 балів** - відтворюється значна частина розглянутої теми. Виявлено знання і розуміння основних положень навчальної дисципліни, проте присутні неточності та/або невідповідності основній темі. Наведені приклади коду частково робочі, проте в загальному відповідають темі; **5 балів** - відстежується загальне розуміння розглянутої теми. Виявлені множинні неточності та невідповідності, пояснення наведеного коду відступні, код функціонує із значними неточностями (або відсутні приклади запуску коду на виконання взагалі);

	3 бали – студент погано розуміє розглянуту тему. Виявлені суттєві неточності та невідповідності. Наведені приклади коду з суттєвими недоліками, або не відповідають темі; **Менше 3 балів** – студент взагалі не розуміє розглянуту тему. Тему не розкрито, кількість викладеного матеріалу не відповідає загальним нормам обраного виду роботи. Наведений код не робочий, або відсутній як такий. **Академічна доброчесність**: Очікується, що лабораторні роботи 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. **Відвідання 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 контрольних робіт	Перелік питань та завдань для проведення підсумкової оцінки знань певних тем до контрольних робіт: 1. Поняття інженерії даних (*Робота інженера по даних. Технології інженера по даних. Навички інженера по даних. Поняття інженера по даних, інженера аналітика, на інженера по знаннях*) 2. Використання Python в інженерії даних (*Основи Python. Засоби Python для роботи з даними. Типи та способи представлення даних. Навички інженера по даних. Поняття інженера по даних, інженера*

	аналітика, на інженера по знаннях) 3. Елементи лінійної алгебри (*Векторний простір. Для чого потрібна лінійна алгебра? Матриці. Лінійна незалежність. Норми. Ортогональна матриця*) 4. Операції над матрицями (*Матриці. Властивості матриць. Операції над матрицями за допомогою Python. Власні значення та власні вектори. Система рекомендацій. Аналіз головних компонентів.*) 5. Основи програмування в Python (*Типи даних. Складні типи даних. Класи та словники. Синтакс мови Python*) 6. Використання Numpy (*Робота з масивами в Numpy, Операції над масивами. Сортування. Доступ до елементів масиву. Основні операції над векторами та тензорами*) 7. Використання Pandas (*Робота з масивами в Pandas. Pandas Dataframe. Операції над масивами даних. Фільтрування та групові операції. Вибірка даних. Візуалізація даних за допомогою Pandas.*) 8. Поняття статистичного аналізу даних (Середнє арифметичне, Медіана, Мода, Середньоквадратичне відхилення, Дисперсія, Математичне очікування, Розподіли даних. Гістограми.) 9. Методи статистичного аналізу даних (*Оцінка розподілу за вибіркою. Ядерна оцінка щільності. Поняття гістограми. Оцінка даних на гістограмі. Розкид випадкової величини*) 10. Методи статистичного аналізу даних (*Центральна гранична теорема. Довірчі інтервали. Передбачувальні інтервали. Перевірка гіпотез.*) 11. Поняття теорії ймовірностей (*Визначення імовірності. Статистичний підхід. Поняття випадковості в теорії ймовірності та статистиці. Операції над подіями. Теорема додавання. Незалежність подій. Умовна імовірність. Теорема Байєса.*) 12. Використання теорії ймовірностей (*Випадкові величини. Функція розподілу. Дискретні випадкові величини. Зв'язок між розподілами. Безперервні величини. Нормальний розподіл. Багатомірний нормальний розподіл*) 13. Особливості візуалізації даних (*Декомпозиція та агрегація даних. Трансформація даних. Поділ даних. Злиття даних. Групування. Сортування. Matplotlib. Фігури. Полотно. Візуалізація.*) 14. Візуалізація засобами Python (*Налаштування фігури. Pandas. Pandas об'єкти. Bar plot. Stacked Bar Plots. Seaborn. Агрегація та сумаризація. Гістограми та розподіл густини. Розкиди або точкові діаграми. Фасетні сітки та категорійні дані.*) 15. Поняття даних (*Життєвий цикл науки про дані. Різновити алгоритмів машинного навчання. Типи та види навчання. Типи даних. Структуровані та неструктуровані дані. Категорійні дані. Часові ряди*) 16. Поняття інженерії ознак (*Поняття оглядового аналізу даних. Засоби для аналізу даних. Інженерія ознак. Поняття ознаки. Трансформація ознак. Масштабування даних.*) 17. Дослідження ознак (*Поняття ознак для структурованих та не структурованих даних. Важливість ознак. Методи визначення важливостей ознак. Використання Scikit-learn для визначення важливості ознак.*) 18. Поняття метрик якості (*Поняття оціночної метрики. Метрики оцінки вирішення задач регресії, класифікації, навчання без учителя,*

	на ін. Метрика точності, акуратності, f1-score.) 19. Оцінки та методи розрахунку метрик (*Розрахунок метрик. Задовільнення критеріїв та оптимізація метрик. Упередженість, якої можна уникнути. Типові помилки даних. Поняття аналізу помилок. Інтерпретація помилки. Інтерпретаційні моделі. Властивості інтерпретаційних моделей.*) 20. Методи скорочення розрядності (*Поняття простору даних. Поняття великої розрядності. Метод аналізу основних компонентів (PCA). Особливості роботи PCA. Поняття Eigenvector. Поняття генів. Дисперсія. Застосування PCA.*) 21. Метод t-SNE для скорочення розрядності (*Вибір системи оцінювання даних. Атрибути якості даних. T-розподілене стохастичне вбудовування сусідів (t-SNE). Масштабування відстані та складність. tSNE проекції. Застосування t-SNE.*) 22. Використання UMAP для скорочення розрядності (*Поняття UMAP. Переваги UMAP. Практичні рекомендації щодо застосування PCA + tSNE/UMAP*) 23. Дослідження даних на основі їх розподілів (*Розподіли та гістограми. Нормальний розподіл Гауса. Інші розподіли. Популяція. Z-нормалізація.*) 24. Колеляція в ознаках даних (*Поняття коваріації. Матриця коваріації. Змінні, що корелюють. Лінійна кореляція. Кореляція Пірсона. Вивчення лінійних залежностей. Метод спроб і помилок. Оптимізація. Кореляція за багатьма змінними. Залишки регресії. Матриця кореляції. Поняття коваріації*) 25. Поняття часових рядів (*Що таке часовий ряд? Що таке графік часових рядів? Моделювання даних часових рядів. Часовий ряд і стаціонарність. Модель наполегливості. Авторегресійна модель. Авторегресійне інтегроване ковзне середнє. Сезонність.*) 26. Аналіз часових рядів (*Аналіз часових рядів. Серійна кореляція, автокореляція. Формування дата-сетів для передбачення аномалій на основі часових рядів.*) 27. Сучасні методи візуалізації даних (*Сучасні досягнення у візуалізації даних. Правильне використання кольорів. Безперевна візуалізація. Показовий рендерінг. Побудова конвеєру візуалізації даних високої розмірності. Колектор навчання. Детальніше про методи: PCA, t-SNE, Umap.*) 28. Розширені можливості візуалізації даних (*Роль машинного навчання у візуалізації даних з високою розмірністю. Алгоритми машинного навчання. Групи алгоритмів та їх призначення. Алгоритми машинного навчання в Scikit-learn.*) 29. Дослідницький аналіз даних і візуалізація за допомогою Python (*Основні поняття та цілі дослідницького аналізу даних (EDA). Роль EDA у процесах науки про дані. Типи даних. Корисні властивості Numerical Python (NumPy) для EDA.*) 30. Дослідницький аналіз даних за допомогою Pandas (*Розширені можливості Pandas. Використання гістограм. Візуалізація від Seaborn. Перекриття екскізів. Емпірична кумулятивна функція розподілу. Теплові карти.*) 31. Поняття системи рекомендацій (*Що таке система рекомендацій? Навіщо потрібні системи рекомендацій? Завантаження даних. Роль дослідницького аналізу даних. Що таке рейтинги? Середній Байєс. Трансформація даних. Обчислення розрідженості матриці.*)

	32. Пошук по схожості для рекомендаційних систем (*Пошук схожих об'єктів. Міри схожості. Спільна фільтрація. Фільтрування на основі вмісту. Робота із набором даних MovieLens.*) 33. Знайомство з PySpark (*Концепція Spark. Поняття RDD. RDD операції. Перегляд на рівні RDD. Керування роботами. Доступні APIs. Ініціалізація PySpark. Створення RDD з PySpark.*) 34. Операції в PySpark (*Операції подій в PySpark. Операції трансформації в PySpark. Функціональне програмування у PySpark. PySpark. flatMap. Pair RDD. Відображення окремих рядків у кілька пар. Поняття MapReduce. MapReduce з PySpark. Об'єднання RDD.*) 35. Spark SQL: Relational Data Processing in Spark (*Труднощі та рішення. Модель Spark. Швидкодія Spark. Оцінка технічного стеку Spark. Поняття Spark SQL. Програмний інтерфейс. Модель даних. Поняття DataFrame. Операції з DataFrame. Запити. Оптимізація та виконання.*) 36. Робота з Spark MLlib (*Генерація коду. Розширення. Розширені функції аналітики. Spark MLlib конвеєри. Дослідницька трансформація. Порядок виконання операцій над даних. Скорочення обчислень. Алгоритми машинного навчання.*) 37. Поняття розподілених систем (*Розподілені системи опрацювання даних. Бази даних. Маштабування баз даних. Поняття маштабування в розподілених системах опрацювання даних.*) 38. Основнні поняття ETL (*Поняття ETL (Extract, Transform, Load). Розбиття даних за ключами. Розбиття даних за файлами. Перекіс даинх. Зсуваюче об'єднання. Інші типи об'єднань. Різниця між ETL та ELT.*) 39. Використання хмарних технологій для побудових конвеєрів даних (*Робочий потік даних в AWS GLU. Azure Data Factory. Серсвіси бізнес даних. Технологічний стек. Потокові дані з Apache Kafka. Побудова Kafka кластеру. Створення та поглинання з Python. Приклади.*) 40. Поняття великих даних (*Великі дані. Особливості великих даних. Алгоритми та методи опрацювання великих даних. Формування дата-сетів на основі сентетичних даних.*)
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

СХЕМА КУРСУ

Тиж.	Тема, план, короткі тези	Форма діяльності	Література. Ресурси в Інтернеті	Завдання, год	Термін виконання, тиж.
1	**Вступ до інженерії даних.** Робота інженера по даних. Технології інженера по даних. Навички інженера по даних. Поняття інженера по даних, інженера аналітика, на інженера по знаннях. Основи Python. Засоби Python для роботи з даними. Типи та способи представлення даних.	лекція	1- 7	2	кінець поточного тижня
	Основи Python та Numpy	лаб. робота	1-5	2	кінець поточного тижня
	Основи програмування в Python	сам. робота	1-5	2,56	кінець поточного тижня
2	**Елементи лінійної алгебри.** Векторний простір. Для чого потрібна лінійна алгебра? Матриці. Лінійна незалежність. Норми. Ортогональна матриця. Власні значення та власні вектори. Система рекомендацій. Аналіз головних компонентів.	лекція	6-9	2	кінець поточного тижня
	Основи роботи з Pandas	лаб. робота	2	2	кінець поточного тижня
	Класи та моделі даних в Python	сам. робота	1-3	2,56	кінець поточного тижня
3	**Методи статистичного аналізу даних.** Оцінка розподілу за вибіркою. Ядерна оцінка щільності. Поняття гістограми. Оцінка даних на гістограмі. Статистичні дані. Середнє значення. Медіана. Мода. Математичне очікування. Дисперсія. Середньоквадратичне відхилення. Вибірка. Властивість рівномірного розподілу. Бімодальний розподіл. Розкид випадкової величини. Центральна гранична теорема. Довірчі інтервали. Передбачувальні інтервали. Перевірка гіпотез.	лекція	1, 6, 10, 11	2	кінець поточного тижня
	Візуалізація даних з Matplotlib	лаб. робота	6	2	кінець поточного тижня
	Опрацювання та підготовка структурованих даних	сам. робота	7-9	2,56	кінець поточного тижня

4	**Теорії ймовірностей у методах опрацювання даних.** Визначення імовірності. Статистичний підхід. Поняття випадковості в теорії ймовірності та статистиці. Операції над подіями. Теорема додавання. Незалежність подій. Умовна імовірність. Теорема Байєса. Випадкові величини. Функція розподілу. Дискретні випадкові величини. Зв'язок між розподілами. Безперервні величини. Нормальний розподіл. Багатомірний нормальний розподіл.	лекція	8, 12, 13	2	кінець поточного тижня
	Засоби візуалізації Seaborn	лаб. робота	2, 6	2	кінець поточного тижня
	Опрацювання та підготовка текстових даних	сам. робота	11	2,56	кінець поточного тижня
5	**Візуалізація даних.** Декомпозиція та агрегація даних. Трансформація даних. Поділ даних. Злиття даних. Групування. Сортування. Matplotlib. Фігури. Полотно. Візуалізація. Налаштування фігури. Pandas. Pandas об'єкти. Bar plot. Stacked Bar Plots. Seaborn. Агрегація та сумаризація. Гістограми та розподіл густини. Розкиди або точкові діаграми. Фасетні сітки та категорійні дані.	лекція	8, 9, 14	2	кінець поточного тижня
	Пошуковий аналіз текстових даних	лаб. робота	8, 12, 14	2	кінець поточного тижня
	Опрацювання та підготовка відео даних	сам. робота	2, 7, 15	2,56	кінець поточного тижня
6	**Дані та ознаки.** Життєвий цикл науки про дані. Різновити алгоритмів машинного навчання. Типи та види навчання. Типи даних. Структуровані та неструктуровані дані. Категорійні дані. Часові ряди. Поняття оглядового аналізу даних. Засоби для аналізу даних. Інженерія ознак. Поняття ознаки. Трансформація ознак. Масштабування даних.	лекція	6, 8, 15	2	кінець поточного тижня
	Опрацювання та перетворення даних	лаб. робота	2, 3, 7	2	кінець поточного тижня
	Опрацювання та підготовка аудіо даних	сам. робота	2, 3, 15	2,56	кінець поточного тижня
7	**Оцінки та метрики.** Поняття оціночної метрики. Метрики оцінки вирішення задач регресії, класифікації, навчання без учителя, на ін. Метрика точності, акуратності, f1-score. Розрахунок	лекція	8, 15 - 20	2	кінець поточного тижня

	метрик. Задовільнення критеріїв та оптимізація метрик. Упередженість, якої можна уникнути. Типові помилки даних. Поняття аналізу помилок. Інтерпретація помилки. Інтерпретаційні моделі. Властивості інтерпретаційних моделей.				
	Описова статистика	лаб. робота	11-13	2	кінець поточного тижня
	Статистичний аналіз даних	сам. робота	11, 12	2,56	кінець поточного тижня
8	**Методи скорочення розрядності.** Поняття простору даних. Поняття великої розрядності. Метод аналізу основних компонентів (PCA). Особливості роботи PCA. Поняття Eigenvector. Поняття генів. Дисперсія. Застосування PCA. Вибір системи оцінювання даних. Атрибути якості даних. T-розподілене стохастичне вбудовування сусідів (t-SNE). Масштабування відстані та складність. tSNE проекції. Застосування t-SNE. Поняття UMAP. Переваги UMAP. Практичні рекомендації щодо застосування PCA + tSNE/UMAP.	лекція	8, 9, 15	2	кінець поточного тижня
	Операції над дата сетами	лаб. робота	3, 4, 9	2	кінець поточного тижня
	Оглядовий аналіз з Pandas	сам. робота	14, 15	2,58	кінець поточного тижня
9	**Кореляція в ознаках даних.** Розподіли та гістограми. Нормальний розподіл Гауса. Інші розподіли. Популяція. Z-нормалізація. Поняття коваріації. Матриця коваріації. Змінні, що корелюють. Лінійна кореляція. Кореляція Пірсона. Вивчення лінійних залежностей. Метод спроб і помилок. Оптимізація. Кореляція за багатьма змінними. Залишки регресії. Матриця кореляції. Поняття коваріації.	лекція	1, 8, 13	2	кінець поточного тижня
	Кореляція даних	лаб. робота	8, 11, 13	2	кінець поточного тижня
	Конвеєр даних з scikit-learn	сам. робота	15	2,56	кінець поточного тижня
10	**Часові ряди.** Що таке часовий ряд? Що таке графік часових рядів? Моделювання даних часових рядів. Часовий ряд і стаціонарність. Модель наполегливості. Авторегресійна модель. Авторегресійне інтегроване ковзне середнє.	лекція	21, 22	2	кінець поточного тижня

	Сезонність. Аналіз часових рядів. Серійна кореляція, автокореляція. Формування дата-сетів для передбачення аномалій на основі часових рядів.				
	Аналіз часових рядів	лаб. робота	12	2	кінець поточного тижня
	Важливість ознак з scikit-learn	сам. робота	15	2,56	кінець поточного тижня
12	**Дослідницький аналіз даних і візуалізація за допомогою Python.** Основні поняття та цілі дослідницького аналізу даних (EDA). Роль EDA у процесах науки про дані. Типи даних. Корисні властивості Numerical Python (NumPy) для EDA. Розширені можливості Pandas. Використання гістограм. Візуалізація від Seaborn. Перекриття екскізів. Емпірична кумулятивна функція розподілу. Теплові карти.	лекція	8, 13	2	кінець поточного тижня
	Побудова конвеєру пошукового аналізу даних	лаб. робота	2, 8, 20	2	кінець поточного тижня
	Аналіз та балансування дата-сетів	сам. робота	22, 23	2,56	кінець поточного тижня
13	**Побудова систем рекомендацій.** Що таке система рекомендацій? Навіщо потрібні системи рекомендацій? Завантаження даних. Роль дослідницького аналізу даних. Що таке рейтинги? Середній Байєс. Трансформація даних. Обчислення розрідженості матриці. Пошук схожих об’єктів. Міри схожості. Спільна фільтрація. Фільтрування на основі вмісту. Робота із набором даних MovieLens.	лекція	1, 20, 23-25	2	кінець поточного тижня
	Використання міри схожості у системах рекомендацій	лаб. робота	8, 14, 22, 23	2	кінець поточного тижня
	Формування дата-сетів на основі сентетичних даних	сам. робота	23	2,56	кінець поточного тижня
14	**Знайомство з PySpark.** Концепція Spark. Поняття RDD. RDD операції. Перегляд на рівні RDD. Керування роботами. Доступні APIs. Ініціалізація PySpark. Створення RDD з PySpark. Операції подій в PySpark. Операції трансформації в PySpark. Функціональне програмування у PySpark. PySpark. flatMap. Pair RDD. Відображення окремих рядків у кілька пар.	лекція	1, 13, 25	2	кінець поточного тижня

	Поняття MapReduce. MapReduce з PySpark. Об’єднання RDD.				
	Програмування ETL логіки	лаб. робота	23-35	2	кінець поточного тижня
	Інтеграція засобів моніторингу із ковеєром даних	сам. робота	23-25	2,56	кінець поточного тижня
15	**Spark SQL: Relational Data Processing in Spark.** Труднощі та рішення. Модель Spark. Швидкодія Spark. Оцінка технічного стеку Spark. Поняття Spark SQL. Програмний інтерфейс. Модель даних. Поняття DataFrame. Операції з DataFrame. Запити. Оптимізація та виконання. Генерація коду. Розширення. Розширені функції аналітики. Spark MLlib конвеєри. Дослідницька трансформація.	лекція	1, 25	2	кінець поточного тижня
	Побудова ETL конвеєрів	лаб. робота	23, 25	4	кінець поточного тижня
	Організація конвеєра даних на основі Apach Kafka	сам. робота	20, 24, 25	2,56	кінець поточного тижня
16	**Маштабовані конвеєри даних.** Розподілені системи опрацювання даних. Поняття маштабування в розподілених системах опрацювання даних. Поняття ETL (Extract, Transform, Load). Розбиття даних за ключами. Розбиття даних за файлами. Перекіс даинх. Зсуваюче об’єднання. Інші типи об’єднань. Різниця між ETL та ELT. Робочий потік даних в AWS GLU. Azure Data Factory. Серсвіси бізнес даних. Технологічний стек. Потокові дані з Apache Kafka. Побудова Kafka кластеру. Створення та поглинання з Python. Приклади. Великі дані. Робота з Великими даними.	лекція	1, 25	2	кінець поточного тижня
	Організація конвеєра даних на основі PySpark Dataframe	сам. робота	20, 23, 25	2,58	кінець поточного тижня