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

Затверджено
На засіданні кафедри СП
Факультету електроніки та
комп'ютерних технологій
Львівського національного
університету імені Івана Франка
(протокол № _ від __ серпня 2021 р.)

Силабус з навчальної дисципліни «Методології дослідження даних», що викладається в межах ОПП "Інженерія програмного забезпечення" (ВПК) першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 121 «Інженерія програмного забезпечення» (ВПК)

Львів 2021

Назва дисципліни	Методології дослідження 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	121 Інженерія програмного забезпечення (ВПК)
Викладачі дисципліни	Ляшкевич Василь Яремович, канд. тех. наук, доцент, доцент
Контактна інформація	vasyl.lyashkevych@lnu.edu.ua, https://electronics.lnu.edu.ua/employee/liashkevych-v-ya
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=309
Інформація про дисципліну	Дисципліна «Методології дослідження даних» є вибіркової дисципліною з блоку "Високопродуктивні технології" спеціальності 121 Інженерія програмного забезпечення для освітньої програми «Високопродуктивний комп'ютинг», яка викладається у 8 семестрі в обсязі 5 кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Навчальну дисципліну розроблено таким чином, щоб надати учасникам необхідні знання, обов'язкові для того, щоб оволодіти базовими поняттями даних, особливості використанням даних, використанням технологій роботи з даними та технології і методології дослідження даних та розв'язувати різні задачі в області науки про дані та систем штучного інтелекту. У дисципліні представлено огляд базових інструментів роботи з даними, знаннями, засобами, які потрібні для вирішення типових завдань при використанні, налаштуванні середовищ та технологій роботи з даними для вирішення проблем в галузі науки про дані.
Мета та цілі дисципліни	**Метою** дисципліни "Методології дослідження даних" є забезпечення майбутніх інженерів навиками дослідження та аналізу даних, оволодіння базовими поняттями, теоретичними знаннями про дані, можливостями інформаційних систем, побудованих на основі опрацювання та аналізу даних, візуалізації даних, побудови конвеєрів для дослідження і перетворення даних з подальшим використанням в різних інноваційних системах при вирішенні задач наук про дані в різних предметних областях

	людської діяльності.
Література для вивчення дисципліни	Основна література: 1. Christopher M. Bishop (2018) Pattern Recognition and Machine Learning, 738p. 2. Sarah Guido (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists, 400p. 3. EMC Education Services (2015) Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 432p. 4. Cole Nussbaumer Knaflic (2015) Storytelling with Data: A Data Visualization Guide for Business Professionals, 288p. 5. Peter Bruce (2017) Statistics for Data Scientists: 50 Essential Concepts, 298p 6. Data Mining: The Complete Guide. – Columbia Engineering, 2023. URL: https://bootcamp.cvn.columbia.edu/blog/data-mining-guide/ 7. Paul Crickard. Data Engineering with Python - Birmingham: Packt Publishing, 2020. - 337 p. - ISBN 978-1-83921-418-9. 8. Wang L., Fu X. Data Mining with Computational Intelligence. –Springer, 2005. –280 p. 9. Wes McKinney. Python for Data Analysis - Sebastopol: O'Reilly Media, 2018. - 522 p. - ISBN 978-1-491-95766-0. 10. Joakim Sundnes. Introduction to Scientific Programming with Python - Lysaker: Simula SpringerBriefs, 2020, Volume 6. - ISBN: 978-3-030-50355-0. (eBook) 11. Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser. Data Structures & Algorithms in Python. Wiley: Courier Westford, 2013. - 748 p. (eBook) 12. Massimo di Pierro. Annotated Algorithms in Python - Chicago: Experts4Solutions, 2017. - 227 p. - ISBN: 978-0-9911604-0-2. 13. Allen B. Downey. Think Stats. Exploratory Data Analysis in Python - Needham: Green Tea Press, 2014. - 244 p. 14. Jake VanderPlas. Python Data Science Handbook - Sebastopol: O`Reilly Media, 2017. - 517 p. - ISBN: 978-1-491-91205-8. 15. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: concepts and techniques - Waltham: Elsevier, 2012. - 703 p. 16. Peter Bruce, Andrew Bruce, Peter Gedeck. Practical Statistics for Data Scientists. - Sebastopol: O`Reilly, 2020. - 329 p. - ISBN: 978-1-492-07294-2. 17. Brian Godsey. Think Like a Data Scientist. - Shelter Island: Manning Publications, 2017. - 299 p. - ISBN: 9781633430273. 18. Meher Krishna Patel. Pandas Guide. - May, 2020. - 62 p. 19. Aurelien Geron. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. - Sebastopol: O`Reilly, 2019. - 482 p. - ISBN: 978-1-492-03264-9. 20. 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 21. 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 22. 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.

	23. Mats Bergdahl, Manfred Ehling, Eva Elvers and others. Handbook on Data Quality Assessment Methods and Tools. - Wiesbaden, 2007. - 139 p. 24. The Ultimate Guide to Basic Data Cleaning: Atlan, 2014. - 66 p. 25. 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. 26. Dimensionality reduction [Режим доступу]: http://bioconductor.org/books/3.15/OSCA.basic/dimensionality-reduction.html 27. Data exploration with alluvial plots [Режим доступу]: https://www.datisticsblog.com/2018/10/intro_easyalluvial/#features 28. Aurélien Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow: O'Reilly, 2017. - 718 p. 29. Rezaul Karim, Mahedi Kaysar. Large Scale Machine Learning with Spark: Packt Publishing, 2016. - 472 p. 30. Andrew Ng. Machine Learning Yarning. - [Електронний ресурс]. - Режим доступу: https://nessie.ilab.sztaki.hu/~kornai/2020/AdvancedMachineLearning/Ng_MachineLearningYearning.pdf 31. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006. 32. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014. 33. Charu C. Aggarwal. Recommender Systems: Springer, 2016. - 518 p. 34. Kishan G. Mehrotra Chilukuri K. Mohan HuaMing Huang. Anomaly Detection Principles and Algorithms: Springer. - 2017. - 229 p. - DOI: https://doi.org/10.1007/978-3-319-67526-8 35. Machine Learning in Computer Vision / N. Sebe, Ira Cohen, Ashutosh Garg, Thomas S. Huang// Springer, 2005. - 249 p. - Режим доступу: http://silverio.net.br/heitor/disciplinas/eeica/papers/Livros/[Sebe]%20-%20Machine%20Learning%20in%20Computer%20Vision.pdf 36. Mark Richards. Software Architecture Patterns. - Sebastopol: O`Reilly Media, 2015. - 45 p. - ISBN: 978-1-491-92424-2. 37. Tomcy John, Pankaj Misra. Data Lake for Enterprises. - Packt Publishing, 2017. - 855p. 38. Viktor Mayer-Schonberger, Kenneth Cukier (2013) Big Data: A Revolution That Will Transform How We Live, Work and Think, 256 p. 39. Alex Holmes. Hadoop in Practice: Manning Publications, 2012. - 537 p. - Режим доступу: https://ia600201.us.archive.org/7/items/HadoopInPractice/Hadoop%20in%20Practice.pdf 40. Apache HBase Team. Apache HBase ™ Reference Guide. - [Електронний ресурс]. - Режим доступу: https://hbase.apache.org/apache_hbase_reference_guide.pdf 41. Google. Cloud Bigtable. - [Електронний ресурс]. - Режим доступу: https://cloud.google.com/bigtable
Обсяг курсу	Кількість кретитів ЄКТС: 5 (150 год), з них: 64 годин аудиторних занять (лекції: 32 год, лабораторні: 32 год.) та 86 год. самостійної роботи.
Очікувані результати	Після завершення цього курсу студент буде:

навчання	- розуміти сутність даних, значення яке вони містять, та критерії їх пошуку; - знати основні принципи розробки алгоритмів та програмного забезпечення розв'язування задач видобування та дослідження даних; - вміти 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. Після вивчення курсу здобувачі набудуть таких компетентностей і програмних результатів: ● **ЗК01.** Здатність до абстрактного мислення, аналізу та синтезу. ● **ЗК02.** Здатність застосовувати знання у практичних ситуаціях. ● **ЗК04.** Здатність спілкуватися іноземною мовою як усно, так і письмово. ● **ЗК05.** Здатність вчитися і оволодівати сучасними знаннями. ● **ЗК06.** Здатність до пошуку, оброблення та аналізу інформації з різних джерел. ● **ФК14.** Здатність брати участь у проектуванні програмного забезпечення, включаючи проведення моделювання (формальний опис) його структури, поведінки та процесів функціонування. ● **ФК16.** Здатність формулювати та забезпечувати вимоги щодо якості програмного забезпечення у відповідності з вимогами замовника, технічним завданням та стандартами. ● **ФК19.** Володіння знаннями про інформаційні моделі даних, здатність створювати програмне забезпечення для зберігання, видобування та опрацювання даних. ● **ФК20.** Здатність застосовувати фундаментальні і міждисциплінарні знання для успішного розв'язання завдань інженерії програмного забезпечення. ● **ФК25.** Здатність обгрунтовано обирати та освоювати інструментарій з розробки та супроводження програмного забезпечення. ● **ФК26.** Здатність до алгоритмічного та логічного мислення. ● **ФК28.** Володіння методами розроблення і впровадження систем підвищеної продуктивності, серверних, мікросервісних, хмаркових, розподілених та інших новітніх технологій. ● **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.

	● **ПРН06.** Вміння вибирати та використовувати відповідну задачі методологію створення програмного забезпечення. ● **ПРН07.** Знати і застосовувати на практиці фундаментальні концепції, парадигми і основні принципи функціонування мовних, інструментальних і обчислювальних засобів інженерії програмного забезпечення. ● **ПРН13.** Знати і застосовувати методи розробки алгоритмів, конструювання програмного забезпечення та структур даних і знань. ● **ПРН18.** Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних. ● **ПРН21.** Знати, аналізувати, вибирати, кваліфіковано застосовувати засоби забезпечення інформаційної безпеки (в тому числі кібербезпеки) і цілісності даних відповідно до розв'язуваних прикладних завдань та створюваних програмних систем. ● **ПРН25.** Вміти застосовувати інноваційні технологічні рішення при розробці високопродуктивних систем. ● **ПРН27.** Знати основи інженерії й аналізу даних та вміти вибрати оптимальні алгоритми і технології для розробки інноваційних рішень при розв'язанні задач наук про дані, вбудованих систем та систем штучного інтелекту.
Ключові слова	Дані, інформація, знання, дослідження даних, розподіли даних, системи дослідження даних, ознаки даних, аналіз даних, системи аналізу даних, системи, що керуються даними.
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 в галузі знань 12 – Інформаційні технології, а також проходження дисциплін: "Архітектури та технології глибинного навчання", "Опрацювання інформації", "Еволюційні, генетичні, евристичні та метаевртаристичні алгоритми".
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

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: 70% семестрової оцінки, максимальна кількість балів 70: - лабораторні роботи: 68,5% оцінки контрольного заміру; максимальна кількість балів 48 (16 лабораторних робіт по 3 бали за кожну). - теоретичний матеріал: 31,5% оцінки контрольного заміру; максимальна кількість балів 22 (2 тести по 9 балів за кожний). • залік: 30% семестрової оцінки, максимально 30 балів. Оцінки за **лабораторні заняття** розподіляються наступним чином: 3 (2 бали за виконання, 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 (1 бали за виконання, 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 (0.5 бали за виконання, 0.5 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 відстуні, 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. **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 даних. *(Формати зберігання даних. Формати представлення даних. Візуалізація даних: Bar chart, Pie chart, Histrogram, Scatter plot, Heatmap, Box plot, Line plot, Violin plot, Bubble plot, 3D plot)* 2. Способи представлення даних. *(Типи наборів даних: записи, хімічні дані, графи. Формати зберігання даних. Формати представлення даних. Візуалізація даних: Bar chart, Pie chart, Histrogram, Scatter plot, Heatmap, Box plot, Line plot, Violin plot, Bubble plot, 3D plot)* 3. Основні поняття структурованих даних. *(Інтерпретація даних. Поняття сутності. Атрибути сутності. Модель сутностей. Модель "Сутність - зв'язок" (ERM). Основні положення про бази даних. Системи управління базами даних. Рівні абстракції даних. Три рівнева архітектура даних. Об'єктно-орієнтовані моделі*

	даних. Класифікація видів даних. Метадані. Категорії метаданих. Ролі та переваги метаданих.) 4. Рівні абстракції даних. *(Рівні абстракції даних. Три рівнева архітектура даних. Об'єктно-орієнтовані моделі даних. Класифікація видів даних. Метадані. Категорії метаданих. Ролі та переваги метаданих.)* 5. Основні положення видобування даних та знань. *(Поняття про процес добування даних, процес KDD (Knowledge Discovery in Databases). Характеристики KDD. Методи та застосування KDD. Дослідження знань в базах даних. Математична база методів добування даних. Статистичні методи. Візуалізація.)* 6. Методи машинного навчання для видобування даних та знань. *(Штучний інтелект. Машинне навчання. Технологія баз даних. Нейронні мережі. Розпізнавання образів. Системи на основі знань. Отримання знань. Пошук інформації. Високопродуктивне обчислення.)* 7. Поняття та представлення даних. *(Статистичні методи. Оцінка розподілу за вибіркою. Емпірична функція розподілу. Статистичні дані. Характеристика розподілу. Точкові оцінки параметрів розподілу. Важливі статистичні дані. Центральна гранична теорема. Квартет Енскомба. Довірчі інтервали. Передбачувальні інтервали. Перевірка гіпотез. Кореляція і коваріація.)* 8. Основи видобування даних. *(Шість стадій опрацювання даних. Причини виникнення та мотивація розвитку методів добування даних. Суть технології добування даних, мультидиципліна рність. Порівняння методів машинного навчання та методів добування даних. Проблеми підготовки даних. Огляд алгоритмів та методів попередньої обробки даних: очистка, перетворення та скорочення.)* 9. Поняття залежності 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.)* 10. Залежність по даних в часі. *(Часові ряди. Що таке графік часових рядів. Моделювання даних часових рядів. Часовий ряд і стаціонарність. Модель наполегливості. Авторегресійна модель. Авторегресійне інтегроване ковзне середнє. Сезонність. Аналіз часових рядів. Серійна кореляція, автокореляція. Формування дата-сетів для передбачення аномалій на основі часових рядів.)* 11. Поняття кореляції та коваріації даних. *(Нормальний розподіл даних. Гістограма. Квартилі. Характеристики нормального розподілу. Поняття популяції. Поняття стандартної помилки середнього значення. Відхилення. Z-нормалізація. Коваріації та Кореляція. Змінні. Лінійна кореляція. Кореляція Пірсона. Метод спраб і помилок. Оптимізація. Регресія.)* 12. Вимірювання та представлення даних. *(Статистичні дані. Розподіли та гістограми. Нормальний розподіл Гауса. Розкид випадкової величини. Розподіли: Біноміальний, Пуасона, Гіпергеометричний, Геометричний, Безперервний, F-розподіл, Chi Squre, T student розподіл, Вейбула та ін. Оцінка розподілу за вибіркою. Симетрія. Моделювання ситуації на основі розподілу даних.)*

	13. Балансування ознак даних. *(Моделювання методом Монте Карло. Вибір бізнес-функції. Моделювання ситуацій. Вибір границі та діапазону даних для вирішення задач. Використання генетичних алгоритмів для балансування ознак даних. Створення популяції. Схрещування. Мутація. Оцінка фітнес-функції. Методи вибору популяції.)* 14. Основні поняття оцінки якості даних. *(Поняття якості даних. Поняття метрик якості. Складність оцінки якості даних. Припущення про дані. Припущення щодо якості даних. Активності щодо якості даних. Ідентифікація якості даних. Моніторинг. Поняття оцінки і вимірювання якості. Контекст якості даних.)* 15. Поняття фреймворку *DQAF*. *(Фреймворк DQAF. Типи вимірювань. Специфічні метрики. Повнота. Позачасовість. Терміновість. Послідовність. Цілісність. Приклади вимірювань. Функції в оцінюванні: збір, обчислення, порівняння, висновки. Періодичний контроль і вимірювання.)* 16. Запис спостережень і висновків про стан якості даних. *(Поняття якості даних. Поняття метрик якості. Контрольний список протоколу аналізу. Лист спостереження. Допоміжні компоненти - призначення та використання, огляд вмісту, визначення термінів.)* 17. Поняття узагальненого аналізу. *(Питання для узагальненого аналізу. Приклади спостережень. Категорії релевантності. Узагальнений аналіз. Фреймворк для Керування якістю даних. Матриця моделі якості.)* 18. Реалізація засобів керування якістю даних. *(Поняття якості даних. Поняття метрик якості. Матриця моделі якості. Вимоги до якості даних. Оцінка вимог. Взаємозв'язок вимог. Вимоги до вмісту даних. Потоки бізнес-процесів. Правила ведення бізнесу. Визначення сутностей і атрибутів. Модель вихідних даних.)* 19. Цільові моделі даних. *(Поняття якості даних. Цільова модель даних. Регіон інтересу для вимірювання якості даних. Дерево рішень. Вимірювання метрик. Внутрішня цілісність даних. Верифікація даних. Кількісні показники. Фактор верифікації. Засоби керування якістю даних. Інформатика.)* 20. Поняття та аналіз предметної області. *(Поняття предметної області. Визначення цілей, задач та границь предметної області. Ситуаційні підходи. Збір та підготовка даних. Тематична модель. Методи аналізу тематичних моделей.)* 21. Математичний опис предметної області. *(Поняття предметної області. Математичний опис ситуацій. Візуалізація предметної області. Приклад тематичної області щодо діагностування комп'ютерних засобів.)* 22. Способи очищення та валідації даних. *(Декомпозиція та агрегація даних. Трансформація даних. Виявлення та вилучення відсутніх даних. Ощищення даних. Інтеграція даних. Скорочення даних. Виявлення аномальних даних. Нормалізація та Стандартизація даних. Кодування даних. Пошук та видалення копій. Валідація даних. Скорочення розмірності. Ітераційні дані.)* 23. Пошуковий аналіз даних. *(Поняття процесу перетворення даних. Інтерпретація даних. Збір та очищення даних. Підготовка даних. Програмні конвеєри даних.*

	Дослідницький аналіз даних. Моделювання даних. Валідація моделей та формування висновків.) 24. Способи і засоби обробки даних. *(Способи подання даних. Візуалізація даних. Повторний цикл аналізу даних. Керування та застосування даних. Поняття штучних даних.)* 25. Накопичення та основи аналізу даних. *(Стратегія накопичення даних. Правила та способи накопичення даних. Вимірювання даних. Кількісний та якісний підходи для накопичення даних. Засоби накопичення даних. Методи участі. Поняття вторинних даних. Огляд даних. Ступені огляду. Порівняння. Пропорції. Рангування. Тренди і зміни. Особливості якісного аналізу даних.)* 26. Методи аналізу основних компонентів. *(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. Практичні рекомендації щодо застосування PCA.)* 27. Методи скорочення розрядності 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 tSNE/UMAP. Методи просторового кодування даних. Робота з даними у векторному просторі.)* 28. Поняття графіки щодо візуалізації даних. *(Маштабована векторна графіка. Поняття координатної площини, фігур, рисунків. Поняття шляхів. Поняття кольору та кольорових гам. Джерела освітлення та камери. 3-D графіки. Конвеєри побудови 3-D графічних об'єктів. Візуалізація баз даних. OLAP. Data Cube. Розмірності. Вимірювання. Поняття агрегації даних. Перетворення даних. Робота з Pandas.)* 29. Засоби Python для візуалізації даних. *(Поняття інформації. Поняття візуалізації даних. Засоби Python для візуалізації даних. Засоби агрегації в Pandas. Візуалізація за допомогою Pandas: Line plot, Bar plot, Stacked Bar Plots. Matplotlib: Simple plot, Figures, Subplots, etc. Seaborn: Tipping percentage, Histogram, Scatter or Point plots, Pair plots, Facet Grids, Box plot, PCA, tSNE, UMAP.)* 30. Поняття штучного інтелекту. *(Життєвий цикл науки про дані. Взаємозв'язок між даними та знаннями. Поняття семантичних мереж. Різновити алгоритмів машинного навчання. Типи та види навчання. Штучний інтелект. Компоненти штучного інтелекту.)* 31. Сучасний штучний інтелект. *(Ера когнітивних обчислень. Вузький штучний інтелект. Когнітивні обчислення. Просторово-часова аналітика. Генеративний інтелект. Супер інтелект. Великі мовні моделі.)* 32. Застосування штучного інтелекту. *(Поняття штучного інтелекту. Типи штучного інтелекту. Спеціалізація ШІ. Компоненти ШІ. Затребувані функції ШІ в індустрії. Хмарні засоби для розроблення інтелектуалізованих систем. Генеративний ШІ. Інтелект як основа великих мовних моделей. Можливості LLM. Конвеєри LLM. Використання баз та сховищ знань. Оптимізація LLM з LoRA. Тренди у великих мовних моделях.)*

	33. Методи колекціонування даних. *(Поняття процесу перетворення даних. Інтерпретація даних. Збір та очищення даних. Підготовка даних. Програмні конвеєри даних. Дослідницький аналіз даних. Моделювання даних. Валідація моделей та формування висновків. Способи подання даних. Візуалізація даних. Повторний цикл аналізу даних. Керування та застосування даних. Поняття штучних даних.)* 34. Методи очищення даних. *(*Декомпозиція та агрегація даних. Трансформація даних. Виявлення та вилучення відсутніх даних. Виявлення аномальних даних. Нормалізація та Стандартизація даних. Кодування даних. Пошук та видалення копій. Валідація даних. Скорочення розмірності. Ітераційні дані.*)* 35. Робота з текстовими даними. *(*Проблеми опрацювання природньої мови, текстових ресурсів. Поняття мовної моделі. Різновиди мовних моделей. Поняття “смислу”. Поняття простору “смислу”. Word Embeddings. Мовні моделі та нейроні мережі.*)* 36. Поняття великих мовних моделей. *(*Велики мовні моделі. Різновид великих мовних моделей. Задачі опрацювання природньої мови в індустрії. Техніка генерації тексту. Створення віршованого тексту. Вступ до конспектування. Генерація заголовків.*)* 37. Знання та соціальні мережі. *(*Поняття соціальної мережі: Наука, Технології, Культура. Концепція соціальної мережі. Популярні соціальні мережі. Історія. Приклади соціальних мереж. Соціальні мережі як технології. Комунікації. Блогери. Керування знаннями. Semantic Web. Аналіз соціальних мереж. Соціальні мереж як джерело знань.*)* 38. Дослідження великих даних. *(*Поняття великих даних. Текстова та мультимедійна інформація. Машина-машина комунікації. Типи великих даних. Труднощі при дослідження великих даних. Hadoop екосистема.*)* 39. Поняття про технології великих даних. *(*Технології великих даних. Компоненти Hadoop екосистеми. Data Spectrum. Spark. Yarn. Основні випадки використання великих даних. Приклади роботи з великими даними.*)* 40. Роль даних у появі нових технологій. *(*Поняття технології. Новітні технології. Визначення зробленого. Життєвий цикл технології. Дані, як каталізатор нових технологій. Цінність даних. Поняття бізнес-даних. Поєднання технологій на основі даних. Доступність даних та технологій. Поняття технологічних трендів. Приклади технологічних трендів на ринку. Поняття упередження. Етичні норми для роботи з даних.*)* 41. Системи дослідження 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 даних. Зсуваюче об’єднання. Інші типи об’єднань. Різниця між ETL та ELT.*)* 42. Поняття ETL. *(Поняття* 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.*)*

	43. Перспективи розвитку дослідження даних. *(*Сучасні тенденції та напрями в дослідженні даних. Поняття інтелекту. Роль машинного навчання для інтелектуальних систем. Складові та характеристики інтелектуальних систем. Збір та збереження знань. Моделі аналізу даних. Моделі генерації знань*.)* 44. Поняття AutoML. *(Поняття AutoML. Звичайні задачі AutoML систем. Відомі AutoML системи. Оптимізація гіперпараметрів. Конвеєр МЛ. Автоматичне машинне навчання. Цикли в AutoML. Semi-AutoML. Автоматична статистика. AutoML визначення. Типи гіперпараметрів. Архітектура конвеєрів. Конвеєр пошуку. Дерево пошуку Монте Карло. Мета-навчання. Мета-дані. Пошук архітектур штучних нейронних мереж. Пошук пошуку. Нейрореволюція. Еволюційний AutoML. Трансфер знань. Трансфер навчання.)* 45. Проектування архітектури даних. *(*Вибір моделі самонавчання. Навчання системи. Оцінка продуктивності. Адаптація та покращення поделей. Безперервний процес навчання інтелектуальних систем. Метрики та методи прийняття рішень. Сучасні засоби для побудови інтелектуальних систем. Архітектура інтелектуальних систем*.)* 46. Проектування систем керованих даних із використанням хмарних технологій. *(Концепція дизайну систем, керованих даними. Архітектура даних. Використання хмарних технологій для проектування систем, керованих даних. Вибір концепції сховища даних: Data Warehouse, Data Lakes, etc. Компоненти хмарної інфраструктури для побудови конвеєрів даних.)* 47. Проектування систем керованих даних. *(Компоненти системи. Поняття конвеєрів обробки даних. Концепція дизайну систем, керованих даними. Архітектура даних. Вибір бази чи сховищ даних. Процеси керування. Синхронізація процесів керування даних. Цілі та гіпотези Вибір стратегії проектування на основі даних.)*
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.
Сертифікація	Сертифікація не є обов’язковим елементом дисципліни, а тільки дозволяє оцінити свої можливості для працевлаштування: - **Python (Basic) Skills** - **Problem Solving (Basic) Skills** - **Problem Solving (Intermediate) Skills** - **Statistics and Machine Learning**

СХЕМА КУРСУ

Тиж.	Тема, план, короткі тези	Форма діяльності	Література. Ресурси в Інтернеті.	Завдання, год	Термін виконання, тиж.
1	**Базові відомості про дані.** Поняття даних. Поняття інформації та знань. Ентропія. 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.	лекція	1-4	2	кінець поточного тижня
	Візуалізація та розподіли даних	лаб. робота	4-8	2	кінець поточного тижня
	Теорія інформації. Теорія кодування. Теорія ймовірностей і статистика. Теорія баз даних. Теорія графів. Алгоритми і обчислювальна складність. Логіка і формальні методи. Формати зберігання даних.	сам. робота	1-4	5.375	кінець поточного тижня
2	**Методи добування 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, мультидиципліна рність. Порівняння методів машинного навчання та методів добування даних. Проблеми підготовки даних. Огляд алгоритмів та методів попередньої обробки даних: очистка, перетворення та скорочення.	лекція	1-4	2	кінець поточного тижня
	Статистичний аналіз даних	лаб. робота	5, 6, 9	2	кінець поточного тижня
	Математичне моделювання. Поняття моделі. Параметри, змінні. Параметризація. Симуляція. Валідація. Тестування. Прогнозування та аналіз.	сам. робота	1, 2	5.375	кінець поточного тижня
3	**Поняття залежностей по даних.** Поняття коваріації. Матриця коваріації. Змінні, що корелюють. Лінійна кореляція. Кореляція Пірсона.	лекція	5-11	2	кінець поточного тижня

	Вивчення лінійних залежностей. Метод спроб і помилок. Оптимізація. Кореляція за багатьма змінними. Залишки регресії. Матриця кореляції. Поняття коваріації. Часові ряди. Що таке графік часових рядів. Моделювання даних часових рядів. Часовий ряд і стаціонарність. Модель наполегливості. Авторегресійна модель. Авторегресійне інтегроване ковзне середнє. Сезонність. Аналіз часових рядів. Серійна кореляція, автокореляція. Формування дата-сетів для передбачення аномалій на основі часових рядів.				
	Засоби покращеного статистичного аналізу	лаб. робота	5, 11-16	2	кінець поточного тижня
	Часові ряди. Декомпозиція. Компоненти часового ряду: тренд, сезонність, нерегулярна складова. Моделі часових рядів. Часові ряди: прогнозування. Прогнозний пакет. Оцінка параметрів для процесів ARMA. Експоненціальне згладжування. Прогнозні комбінації.	сам. робота	8, 9, 12	5.375	кінець поточного тижня
4	**Вимірювання та балансування ознак даних.** Статистичні дані. Розподіли та гістограми. Нормальний розподіл Гауса. Інші розподіли. Оцінка розподілу за вибіркою. Ядерна оцінка щільності. Властивість рівномірного розподілу. Бімодальний розподіл. Розкид випадкової величини. Центральна гранична теорема. Довірчі інтервали. Передбачувальні інтервали. Перевірка гіпотез. Популяція. Z-нормалізація. Моделювання методом Монте Карло. Вибір границі та діапазону даних для вирішення задач. Використання генетичних алгоритмів для балансування ознак даних.	лекція	12-21	2	кінець поточного тижня
	Поняття метрик якості даних	лаб. робота	19-23	2	кінець поточного тижня
	Якість даних. Моделі якості даних. Метрики якості даних. Використання метрик якості даних в конвеєрах даних.	сам. робота	21-23	5.375	кінець поточного тижня

5	**Метрики якості даних.** Поняття оціночної метрики. Метрики оцінки вирішення задач регресії, класифікації, навчання без учителя, на ін. Метрика точності, акуратності, f1-score. Розрахунок метрик. Задовільнення критеріїв та оптимізація метрик. Упередженість, якої можна уникнути. Типові помилки даних. Поняття аналізу помилок. Інтерпретація помилки. Інтерпретаційні моделі. Властивості інтерпретаційних моделей. Бізнес-метрики. Інтеграція бізнес-метрик у конвеєрі даних.	лекція	18-23	2	кінець поточного тижня
	Засоби моніторингу якості даних	лаб. робота	19, 22-24	2	кінець поточного тижня
	Поняття оціночної метрики. Бізнес-метрики. Виявлення проблем з даними. Поняття аналізу помилок. Оцінювання ресурсів даних в індустрії. Оцінювання ресурсів даних для наукових експериментів. Керування якістю даних в Informatica.	сам. робота	19, 22	5.375	кінець поточного тижня
6	**Моделювання предметної області.** Поняття предметної області. Визначення цілей, задач та границь предметної області. Ситуаційні підходи. Збір та підготовка даних. Тематична модель. Методи аналізу тематичних моделей. Візуалізація предметної області. Оцінка та покарщення моделей. Формалізований та неформалізований опис предметної області. Текстовий опис предметної області. Використання великих мовних моделей для опису предметної області. Ієрархічний підхід для опису предметної області.	лекція	3-4, 8, 13-14	2	кінець поточного тижня
	Класифікаційний аналіз	лаб. робота	1-4	2	кінець поточного тижня
	NoSQL бази даних. Моделі сутностей. Схеми даних. RDF сховища. Векторні бази даних. Пошук інформаційї за подібністю. Можливість використання індексації. Застосування векторних баз даних.	сам. робота	24, 25, 27	5.375	кінець поточного тижня

7	**Методи видобування знань.** Методи видобування знаня. Виявлення прихованої інформації. Методи прийтяння рішення. Кластеризація. Класифікація. Регресія. Асоціативні правила. Аналіз аномалій. Аналіз текстових ресурсів. Аналіз зображень. Аналіз соціальних мереж. Дерева рішень. Ансамблі моделей. Зв'язкові правила. Методи глибинного навчання.	лекція	9-10, 15, 17, 19	2	кінець поточного тижня
	Кластерний аналіз	лаб. робота	1-4	2	кінець поточного тижня
	Методи видобування знань. Кластерний аналіз. Регресійний аналіз. Класифікація. Асоціативні правила та дерева рішець. Мовні моделі.	сам. робота	6, 8, 15	5.375	кінець поточного тижня
8	**Методи очищення даних та знань.** Декомпозиція та агрегація даних. Трансформація даних. Виявлення та вилучення відсутніх даних. Виявлення аномальних даних. Нормалізація та Стандартизація даних. Кодування даних. Пошук та видалення копій. Валідація даних. Скорочення розмірності. Ітераційні дані. Таксономія. Поняття таксономії тематичної області. Діаграма Хассе. Експертні системи. Системи логічного висновку. Очищення правил-продукцій. Приклади та контр-приклади в базах знань.	лекція	14, 17, 20, 24	2	кінець поточного тижня
	Методи очищення даних	лаб. робота	24	2	кінець поточного тижня
	Візуальний аналіз даних. Використання кольорових гам. Типи графіків та діаграм. Сучасні 3D засоби візуалізації даних. Масштабування розмірності даних.	сам. робота	25-30	5.375	кінець поточного тижня
9	**Основні положення аналізу даних.** Поняття процесу перетворення даних. Інтерпретація даних. Збір та очищення даних. Підготовка даних. Програмні конвеєри даних. Дослідницький аналіз даних. Моделювання даних. Валідація моделей та формування висновків. Способи подання даних. Візуалізація даних. Повторний цикл аналізу даних.	лекція	1-4, 13, 16	2	кінець поточного тижня

	Керування та застосування даних. Поняття штучних даних.				
	Знайомство з Apache Spark	лаб. робота	1-4, 29	2	кінець поточного тижня
	Web-scraping. Пошук і колекціонування даних. Маркетингові технології колекціонування даних. Цілі та критерії для пошуку даних. Метрики якості для пошуку даних. Критерії вибору даних.	сам. робота	23-25, 29	5.375	кінець поточного тижня
10	**Особливості візуального аналізу даних.** 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. Методи просторового кодування даних. Робота з даними у векторному просторі.	лекція	4, 25-27	2	кінець поточного тижня
	Робота Spark-ML	лаб. робота	1-4, 29	2	кінець поточного тижня
	Методи очищення даних. Кодування категоріальних змінних. Нормалізація та стандартизація даних. Пошук аномалій. Методи валіації та забезпечення якості даних.	сам. робота	24	5.375	кінець поточного тижня
11	**Дані і штучний інтелект.** Життєвий цикл науки про дані. Взаємозв'язок між даними та знаннями. Поняття семантичних мереж. Різновити алгоритмів машинного навчання. Типи та види навчання. Штучний інтелект. Компоненти штучного інтелекту. Ера когнітивних обчислень. Вузький штучний інтелект. Когнітивні обчислення. Просторово-часова аналітика. Генеративний інтелект. Супер інтелект. Великі мовні моделі.	лекція	1-2, 15, 19, 28, 30	2	кінець поточного тижня

	Покращені методи дослідження даних	лаб. робота	30-33	2	кінець поточного тижня
	Поняття структурованих та неструктурованих даних. Поняття текстових ресурсів. Принципи опрацювання природньої мови у текстових ресурсах. Сутності та їх властивості. Мовні моделі. Мовні моделі на schema.org.	сам. робота	30-31, 34-35	5.375	кінець поточного тижня
12	**Бази знань та онтології.** Декомпозиція та агрегація даних. Трансформація даних. Поділ даних. Визначення факту, правила, поняття, події та взаємозв'язків між ними. Поняття словника. Принципи збереження знань. Сховище знань. Експертні системи. Керування знаннями. Поняття онтології. Семантична модель онтології. Графи знань. Формат представлення знань. Контекст. Виразність. Семантичний рівень. Коцептуальний рівень. Абстрактний рівень.	лекція	28, 30-32	2	кінець поточного тижня
	Видобування даних за допомогою NLP	лаб. робота	8, 12-15	2	кінець поточного тижня
	Використання баз знань в індустрії та при виконанні наукових досліджень. Способи побудови баз та сховищ знань. Семантичні зв'язки між даними. Онтології. Соціальні мережі як технології. Соціальні мережі як бази даних та знань.	сам. робота	30, 33, 34	5.375	кінець поточного тижня
13	**Самонавчання та інтелектуальні системи.** Поняття інтелектуалізованих та інтелектуальних систем. Поняття інтелекту. Роль машинного навчання для інтелектуальних систем. Складові та характеристики інтелектуальних систем. Збір та збереження знань. Моделі аналізу даних. Моделі генерації знань. Вибір моделі самонавчання. Навчання системи. Оцінка продуктивності. Адаптація та покращення поделей. Безперервний процес навчання інтелектуальних систем. Метрики та методи прийняття рішень. Сучасні засоби для побудови	лекція	32-35	2	кінець поточного тижня

	інтелектуальних систем. Архітектура інтелектуальних систем.				
	Регресійний та дискримінантний способи дослідження даних	лаб. робота	1-5, 32-35	2	кінець поточного тижня
	Великі дані. Ресурси великих даних. Переваги великих даних. Розвиток бізнесу. Маркетинг та продажі. Проблеми з великими даними. Використання великих даних для наукових досліджень.	сам. робота	36-40	5.375	кінець поточного тижня
14	**Проектування систем, керованих даними.** Визначення сфери інтересу. Визначення цілей. Вибір метрик. Вибір ресурсів та колекцій даних. Візуалізація даних. Визначення алгоритму прийняття рішень. Концепція дизайну систем, керованих даними. Архітектура даних. Вибір бази чи сховищ даних. Процеси керування. Синхронізація процесів керування даних. Цілі та гіпотези Вибір стратегії проектування на основі даних. Використання хмарних технологій для проектування систем, керованих даних.	лекція	36-38	2	кінець поточного тижня
	Видобування даних за допомогою часових рядів	лаб. робота	40-41	2	кінець поточного тижня
	Дані та інновації. Використання даних у різних галузях. Економічний вплив даних. Дані та суспільство. Етичні питання в обробці та використанні даних. Захист даних та кібербезпека. Майбутнє дослідження даних.	сам. робота	30, 40	5.375	кінець поточного тижня
15	Концепція великих даних. 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	39-43	2	кінець поточного тижня
	Комбіновані методи дослідження даних	лаб. робота	36-39	2	кінець поточного тижня

	Підготовлення даних до аналізу. Побудова конвеєрів даних. Валідація та подання даних. Зберезення даних. Методи та підходи до аналізу даних. Конвеєри дослідження даних.	сам. робота	36-38	5.375	кінець поточного тижня
16	**Проектування інтелектуалізованих маштабованих розподілених систем.** 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 даних. 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.	лекція	42-43	2	кінець поточного тижня
	Дослідження соціальних мереж	лаб. робота	39-41	2	кінець поточного тижня
	Поняття інтелектуальних систем. Розподілені технології. Поняття паралелізму. Розподілені бази та сховища даних. Організація кластерів даних. Моделі аналізу даних. Організація обчислювальних кластерів. Архітектура розподілених систем.	сам. робота	37, 39, 41	5.375	кінець поточного тижня