

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Луцький національний технічний університет
Факультет комп'ютерних та інформаційних технологій
Кафедра комп'ютерних наук

DATA SCIENCE

Методичні вказівки до самостійної роботи

з навчальної дисципліни для здобувачів освітньої програми «Комп'ютерні науки»
галузі знань F Інформаційні технології спеціальності F3 Комп'ютерні науки денної
та заочної форм навчання

Луцьк

2025

УДК 004.94:004.8(07)

Д22

До друку

Голова вченої ради ФКІТ _____ І. С. Кондіус
(підпис)

Електронна копія друкованого видання передана для внесення в репозитарій ЛНТУ
директор бібліотеки _____ Н.П. Поліщук
(підпис)

Затверджено вченою радою ФКІТ,
протокол №__ від " __ " _____ 2025 року.

Розглянуто та схвалено на засіданні кафедри комп'ютерних наук ЛНТУ,
протокол №__ від " __ " _____ 2025 року.

Завідувач кафедри КН _____ В. О. Ліщина

Укладач: В. О. Ліщина, к.т.н., доцент, завідувач кафедри комп'ютерних наук ЛНТУ

(підпис)

Рецензент: Н. М. Ліщина, к.т.н., доцент, завідувач кафедри інженерії програмного забезпечення ЛНТУ

(підпис)

Data Science: методичні вказівки до виконання самостійної роботи для здобувачів першого (бакалаврського) рівня освітньо-професійної програми "Комп'ютерні науки" галузі знань F Інформаційні технології спеціальності F3 Комп'ютерні науки денної та заочної форм навчання / уклад. В.О. Ліщина. Луцьк: ЛНТУ, 2025. 24 с.

У методичних вказівках наведений перелік питань, що охоплюють зміст дисципліни, запропоновано завдання для самостійного опрацювання, завдання для індивідуальних завдань, описано порядок поточного і підсумкового оцінювання знань з дисципліни.

Призначені для студентів спеціальності F3 "Комп'ютерні науки" денної та заочної форм навчання.

ЗМІСТ

ВСТУП.....	4
САМОСТІЙНА РОБОТА СТУДЕНТІВ	6
КОНТРОЛЬНІ ЗАПИТАННЯ ДЛЯ САМОДІАГНОСТИКИ.....	7
ІНДИВІДУАЛЬНА РОБОТА.....	10
ПОРЯДОК ПОТОЧНОГО І ПІДСУМКОВОГО ОЦІНЮВАННЯ ЗНАНЬ З ДИСЦИПЛІНИ	11
ОРІЄНТОВНІ ТЕСТОВІ ЗАВДАННЯ	15
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	21

ВСТУП

Головною метою навчальної дисципліни «Data Science» є формування у студентів фундаментальних знань і практичних навичок у сфері збору, обробки, аналізу, інтерактивної візуалізації та інтерпретації великих обсягів даних для прийняття обґрунтованих рішень у різних сферах людської діяльності. Дисципліна спрямована на розуміння ключових концепцій інженерії даних, прикладного машинного навчання та використання технологій штучного інтелекту, зокрема генеративного ШІ (на базі великих мовних моделей, таких як ChatGPT), для оптимізації процесів та підвищення ефективності розв'язання практичних завдань.

Завдання вивчення курсу «Data Science» полягає в отриманні здобувачами теоретичної та практичної підготовки з методів інтелектуального аналізу даних; набутті практичних навичок роботи з реляційними базами даних за допомогою мови SQL; вивченні інструментів обробки та завантаження даних (pandas) у мові програмування Python; освоєнні сучасних технологій візуалізації та побудови аналітичних дашбордів; а також ознайомленні з процесом навчання і валідації моделей машинного навчання.

Вивчення дисципліни забезпечує формування таких компетентностей для спеціальності «Комп'ютерні науки»:

Інтегральна компетентність: здатність розв'язувати складні спеціалізовані задачі та практичні проблеми у галузі комп'ютерних наук, що характеризуються комплексністю та невизначеністю умов, із застосуванням теорій та методів інформаційних технологій.

Загальні компетентності:

К 01. Здатність до абстрактного мислення, аналізу та синтезу.

К 02. Здатність застосовувати знання у практичних ситуаціях.

К 05. Здатність вчитися і оволодівати сучасними знаннями.

Спеціальні (фахові) компетентності:

К 12. Здатність продемонструвати знання та вміння використовувати сучасні методи інтелектуального аналізу даних та машинного навчання.

К 15. Здатність проектувати та розробляти програмне забезпечення для обробки, фільтрації та аналізу великих масивів інформації.

К 20. Здатність застосовувати фундаментальні знання для успішного розв'язання завдань аналізу даних та бізнес-аналітики.

Результати навчання. Після завершення вивчення дисципліни здобувач повинен:

ПР 02. Знати і застосовувати сучасні методи доменного, системного аналізу даних та математичного моделювання.

ПР 08. Проводити аналіз предметної області, розробляти моделі даних, проектувати бази даних та створювати SQL-запити середньої складності.

ПР 11. Вміти використовувати мову Python та спеціалізовані бібліотеки (NumPy, pandas, Matplotlib, Seaborn, Scikit-learn) для очищення, візуалізації даних та побудови прогнозних моделей.

ПР 14. Володіти навичками інтеграції генеративного ШІ (наприклад, ChatGPT) в аналітичні робочі процеси для генерації коду, підготовки звітів та формування бізнес-рекомендацій.

САМОСТІЙНА РОБОТА СТУДЕНТІВ

Самостійна робота є основним видом засвоєння навчального матеріалу у позааудиторний час. Вона включає вивчення лекційного матеріалу, опрацювання рекомендованої літератури, підготовку до лабораторних занять та виконання комплексного індивідуального завдання (КПЗ). Особливе місце займає робота з інтерактивними навчальними курсами на платформах дистанційної освіти (DataCamp, Prometheus, Coursera).

Перелік тем для самостійного опрацювання:

Тема 1. Вступ до аналізу даних та концепцій штучного інтелекту.

Теоретичні питання для опрацювання: Предметна область Data Science, етапи аналізу даних (CRISP-DM), основи Data Literacy, концепція генеративного штучного інтелекту та великих мовних моделей (LLM), використання ChatGPT в аналітиці.

Запитання для самоконтролю:

- 1) Що таке Data Science та які її основні компоненти?
- 2) Опишіть життєвий цикл проекту з аналізу даних.
- 3) Які етичні аспекти використання генеративного ШІ в аналізі даних?
- 4) Як ChatGPT може бути корисним для розробки коду аналітика?

Рекомендовані онлайн-курси: Prometheus «Аналіз даних та статистичне виведення на мові R», Coursera «Introduction to Data Science» (IBM), DataCamp «Introduction to Data», «Introduction to ChatGPT».

Тема 2. Робота з реляційними базами даних та мовою SQL.

Теоретичні питання для опрацювання: Поняття реляційних баз даних, структура таблиць, типи даних. Написання SQL-запитів: фільтрація (WHERE), сортування (ORDER BY), групування (GROUP BY), об'єднання таблиць (JOIN).

Запитання для самоконтролю:

- 1) Опишіть основні типи JOIN-операцій у SQL.

- 2) Яка різниця між операторами WHERE та HAVING?
- 3) Як працюють агрегатні функції у реляційних базах даних?
- 4) Що таке первинний та зовнішній ключі?

Рекомендовані онлайн-курси: Coursera «Data Science Fundamentals with Python and SQL» (IBM), DataCamp «Introduction to SQL», «Intermediate SQL».

Тема 3. Інженерія даних та автоматизація імпорту в Python.

Теоретичні питання для опрацювання: Поняття Data Engineering, архітектура ETL/ELT. Бібліотека pandas для Python: читання та запис файлів (CSV, Excel, JSON), підключення до SQL баз даних за допомогою SQLAlchemy, обробка пропущених значень.

Запитання для самоконтролю:

- 1) Що таке процес ETL та які його основні етапи?
- 2) Як зчитати CSV-файл та SQL-таблицю у DataFrame бібліотеки pandas?
- 3) Опишіть методи обробки пропущених значень у наборах даних.
- 4) Які переваги використання бібліотеки pandas над стандартними списками Python?

Рекомендовані онлайн-курси: Prometheus «Обробка та аналіз великих даних», DataCamp «Introduction to Data Engineering», «Streamlined Data Ingestion with pandas»

Тема 4. Аналітична візуалізація та розвідувальний аналіз даних (EDA).

Теоретичні питання для опрацювання: Мета розвідувального аналізу даних, типи візуалізацій (гістограми, діаграми розсіювання, ящики з вусами). Бібліотеки Matplotlib та Seaborn. Створення інтерактивних дашбордів (Plotly, Dash).

Запитання для самоконтролю:

- 1) Яка роль візуалізації у розвідувальному аналізі даних?
- 2) Опишіть структуру бібліотеки Matplotlib.
- 3) Чим Seaborn відрізняється від Matplotlib?
- 4) Що таке дашборд та які принципи його проектування?

Рекомендовані онлайн-курси: Prometheus «Візуалізація даних», Coursera «Data Visualization with Python» (IBM), DataCamp «Introduction to Data Science in Python».

Тема 5. Основи машинного навчання та прогнозного моделювання.

Теоретичні питання для опрацювання: Класифікація задач машинного навчання (навчання з учителем / без учителя). Лінійна та логістична регресія. Метрики якості моделей (MSE, R2, Accuracy, Precision, Recall, F1). Побудова моделей у Scikit-learn.

Запитання для самоконтролю:

- 1) Яка різниця між навчанням з учителем та без учителя?
- 2) Як працює модель лінійної регресії?
- 3) Які метрики використовуються для оцінки моделей класифікації?
- 4) Що таке перенавчання (overfitting) моделі?

Рекомендовані онлайн-курси: Prometheus «Машинне навчання», Coursera «Machine Learning» (Stanford), DataCamp «Machine Learning with Python».

Тема 6. Сторітелінг на основі даних та презентація результатів.

Теоретичні питання для опрацювання: Принципи сторітелінгу в аналітиці, підготовка бізнес-звітів, дизайн презентацій результатів аналізу. Використання інструментів Power BI та Tableau.

Запитання для самоконтролю:

- 1) У чому полягає суть сторітелінгу в аналізі даних?
- 2) Які основні правила вибору правильного типу діаграми для бізнес-користувача?
- 3) Опишіть етапи створення аналітичного звіту в Tableau.
- 4) Як правильно формулювати рекомендації на основі даних?

Рекомендовані онлайн-курси: Coursera «Data Visualization and Communication with Tableau» (Duke), DataCamp «Storytelling with Data».

КОНТРОЛЬНІ ЗАПИТАННЯ ДЛЯ САМОДІАГНОСТИКИ

1. Що таке Data Science та які її основні міждисциплінарні зв'язки?
2. Які етапи включає життєвий цикл проекту за методологією CRISP-DM?
3. У чому різниця між поняттями Data Science, Data Mining та Big Data?
4. Що таке Data Literacy (грамотність щодо даних) та чому вона важлива для фахівців з комп'ютерних наук?
5. Які основні джерела даних використовуються в сучасній аналітиці?
6. Що таке структуровані, напівструктуровані та неструктуровані дані?
7. Яка роль генеративного штучного інтелекту (Generative AI) у сучасному аналізі даних?
8. Що таке великі мовні моделі (LLM) та як вони навчаються?
9. Як використовувати ChatGPT для генерації та налагодження коду аналізу даних на Python?
10. У чому полягають обмеження та ризики використання великих мовних моделей (наприклад, галюцинації)?
11. Які етичні вимоги висуваються до обробки персональних даних у проектах Data Science?
12. Що таке оперативна розробка (prompt engineering) та які її основні техніки?
13. Як формулювати ефективні запити до ChatGPT для отримання аналітичних інсайтів?
14. Які бібліотеки Python є базовими для наукових обчислень (NumPy, SciPy)?
15. Що таке репрезентативність вибірки даних?
16. Як виявити та обробити викиди (outliers) у наборі даних?
17. Які існують підходи до нормалізації та стандартизації ознак?

18. Що таке реляційна база даних та реляційна таблиця?
19. Яка роль СУБД та мови SQL в архітектурі Data Science проектів?
20. Опишіть синтаксис та порядок виконання операторів у SQL SELECT запиті.
21. Як працює фільтрація даних за допомогою оператора WHERE?
22. Які логічні та порівняльні оператори використовуються в SQL WHERE?
23. Як виконується сортування результуючого набору за допомогою ORDER BY?
24. Опишіть роботу агрегатних функцій (SUM, AVG, COUNT, MIN, MAX) у SQL.
25. Як працює оператор GROUP BY та для чого він використовується?
26. У чому полягає різниця між фільтрацією у WHERE та HAVING?
27. Які типи зв'язків між таблицями існують у реляційних базах даних?
28. Опишіть роботу оператора INNER JOIN.
29. Чим відрізняються LEFT JOIN, RIGHT JOIN та FULL JOIN?
30. Що таке CROSS JOIN та self-join і коли вони застосовуються?
31. Як працюють підзапити (subqueries) у SQL і де їх можна використовувати?
32. Опишіть призначення та синтаксис операторів UNION та UNION ALL.
33. Що таке віконні функції (window functions) в SQL та чим вони відрізняються від GROUP BY?
34. Які особливості використання SQLite у локальних аналітичних проектах?
35. Що таке інженерія даних (Data Engineering) та яка її мета?
36. Яка різниця між архітектурними підходами ETL та ELT?
37. Що таке конвеєр даних (data pipeline) та які основні інструменти його автоматизації?

38. Опишіть структуру DataFrame та Series у бібліотеці pandas.
39. Як завантажити дані з CSV та Excel файлів у DataFrame за допомогою pandas?
40. Які методи pandas використовуються для попереднього перегляду даних (head, info, describe)?
41. Як виконується вибірка (indexing) даних у DataFrame за допомогою loc та iloc?
42. Опишіть підходи до фільтрації рядків DataFrame за логічними умовами.
43. Як працює метод groupby() у pandas та які аналогії він має з SQL?
44. Які методи використовуються для злиття та об'єднання DataFrame (merge, concat, join)?
45. Як обробити пропущені значення (NaN) у pandas за допомогою fillna() та dropna()?
46. Що таке типізація даних у pandas та як змінити тип стовпця (astype)?
47. Як працює векторна обробка даних у pandas та NumPy?
48. Які методи pandas дозволяють ефективно зчитувати дані великого обсягу порціями (chunksize)?
49. Як імпортувати дані з бази даних directly у pandas DataFrame за допомогою read_sql()?
50. Що таке парсинг веб-сторінок (web scraping) та які бібліотеки Python для цього використовуються?
51. Що таке розвідувальний аналіз даних (Exploratory Data Analysis, EDA) та які його задачі?
52. Які основні типи візуалізації використовуються для аналізу розподілу однієї змінної?
53. Що таке гістограма та як правильно вибрати кількість інтервалів (bins)?

54. Яку інформацію надає діаграма «ящик з вусами» (boxplot)?
55. Які типи графіків підходять для візуалізації взаємозв'язку між двома числовими змінними?
56. Що таке діаграма розсіювання (scatter plot) та як вона допомагає виявити кореляцію?
57. Опишіть можливості та структуру бібліотеки Matplotlib.
58. Чим бібліотека Seaborn відрізняється від Matplotlib і які її ключові переваги?
59. Як побудувати матрицю кореляції та візуалізувати її у вигляді теплової карти (heatmap)?
60. Які інструменти використовуються для створення інтерактивних графіків (Plotly)?
61. Що таке аналітичний дашборд та які ключові елементи він містить?
62. Опишіть архітектуру бібліотеки Dash для створення веб-додатків аналізу даних на Python.
63. Які основні правила дизайну дашбордів для бізнес-користувачів?
64. Які існують підходи до візуалізації часових рядів?
65. Як візуалізувати категоріальні дані за допомогою стовпчастих діаграм (bar plots)?
66. Що таке мультипанельні графіки (Faceting) та як їх створювати в Seaborn?
67. Як підготувати інтерактивний звіт у системах бізнес-аналітики Power BI або Tableau?
68. Що таке машинне навчання (Machine Learning) та які його основні класи?
69. У чому різниця між навчанням з учителем (Supervised) та без учителя (Unsupervised)?
70. Які задачі вирішує регресійний аналіз?

71. Опишіть математичну суть моделі лінійної регресії.
72. Що таке функція втрат та як метод найменших квадратів використовується для навчання регресії?
73. Які існують метрики оцінки якості регресійних моделей (MSE, MAE, RMSE, R-squared)?
74. Що показує коефіцієнт детермінації R^2 ?
75. Які задачі вирішує класифікація в машинному навчанні?
76. Опишіть роботу моделі логістичної регресії для бінарної класифікації.
77. Що таке матриця помилок (confusion matrix) класифікації?
78. Дайте визначення метрикам Accuracy, Precision, Recall.
79. Що таке F1-міра та чому вона важлива при незбалансованих класах?
80. Які основні кроки підготовки даних перед тренуванням моделей ML (Train/Test Split)?
81. Що таке перенавчання (overfitting) та недонавчання (underfitting) моделі?
82. Як метод крос-валідації допомагає оцінити узагальнюючу здатність моделі?
83. Опишіть роботу бібліотеки Scikit-learn для створення конвеєрів моделей (Pipeline).
84. Як працює метод найближчих сусідів (KNN) у задачах класифікації?
85. Що таке дерева рішень (Decision Trees) та як вони будуються?
86. У чому полягає ідея ансамблевих методів (наприклад, Random Forest)?
87. Опишіть задачу кластеризації та алгоритм К-середніх (K-Means).
88. Що таке зменшення розмірності даних та для чого використовується метод головних компонент (PCA)?
89. Що таке сторітелінг на основі даних (Data Storytelling) та чому він важливий?

90. Які основні етапи побудови нарративу навколо аналітичних результатів?
91. Як адаптувати презентацію технічних результатів Data Science для менеджменту компанії?
92. Яку структуру повинен мати фінальний аналітичний звіт проекту?
93. Як обрати правильний тип візуалізації залежно від аналітичного повідомлення?
94. Що таке KPI (ключові показники ефективності) та як вони визначаються на основі даних?
95. Опишіть роль метрик у бізнес-аналітиці.
96. Які підходи використовуються для аналізу сезонності та трендів у продажах?
97. Як оцінити економічний ефект від впровадження моделі машинного навчання?
98. Що таке A/B тестування та як аналізуються його результати?
99. Як забезпечити відтворюваність (reproducibility) досліджень у Data Science?
100. Опишіть роль Jupyter Notebook як інструменту для проведення та документування досліджень.

ІНДИВІДУАЛЬНА РОБОТА

Комплексне практичне індивідуальне завдання (КПЗ) є формою перевірки вміння здобувача самостійно застосовувати теоретичні знання для розв'язання прикладних задач з аналізу даних. Виконання та захист КПЗ є обов'язковою умовою для отримання позитивної підсумкової оцінки з дисципліни.

Тема КПЗ: «Аналітичне дослідження та моделювання набору даних з обраної предметної області»

Завдання КПЗ передбачає виконання таких етапів:

1. Вибір предметної області та набору даних (dataset): Студент самостійно обирає реальний набір даних (наприклад, з платформи Kaggle або відкритих джерел даних) обсягом не менше 1000 рядків та 10 ознак.

2. Інженерія даних: Завантаження даних з файлу (CSV/Excel) у СУБД SQLite. Написання 3-4 SQL-запитів для фільтрації, агрегації та групування даних. Імпорт результатів SQL-запитів у Python за допомогою pandas.

3. Розвідувальний аналіз (EDA) та візуалізація: Опис структури даних, типів змінних. Обробка пропусків та аномальних значень. Побудова не менше 5 інформативних графіків (розподіли ознак, кореляційна матриця, діаграми розсіювання, лінійні графіки) за допомогою Matplotlib та Seaborn.

4. Прогнозно-аналітичне моделювання: Побудова та навчання моделі машинного навчання (лінійна або логістична регресія) для прогнозування цільової ознаки. Розбиття на вибірки, оцінка якості моделі на основі відповідних метрик (MSE, R2 або Accuracy, F1-score).

5. Створення аналітичного звіту (сторітелінг): Оформлення результатів у вигляді Jupyter Notebook з текстовими поясненнями та висновками. Розробка презентації (до 10 слайдів) або інтерактивного дашборду (в Tableau, Power BI або Dash) для представлення ключових результатів бізнес-користувачам.

Вимоги до оформлення:

Звіт з КПЗ має бути представлений у вигляді Jupyter Notebook (файл .ipynb) та короткої пояснювальної записки (до 10 сторінок у форматі .docx) з описом предметної області, результатів аналізу та побудованих моделей. Захист КПЗ проводиться у формі презентації результатів перед аудиторією або викладачем.

ПОРЯДОК ПОТОЧНОГО І ПІДСУМКОВОГО ОЦІНЮВАННЯ ЗНАНЬ З ДИСЦИПЛІНИ

Оцінювання знань здобувачів вищої освіти здійснюється за 100-бальною накопичувальною шкалою ЄКТС та національною 4-бальною шкалою.

Підсумковий бал з дисципліни визначається як середнє арифметичне балів залікових модулів, кожен з яких оцінюється від 0 до 100 балів. Протягом семестру здобувачі виконують 3 залікові модулі, які охоплюють лабораторні роботи, модульні контрольні роботи (тестування), самостійну роботу та комплексне індивідуальне завдання (КПЗ).

Таблиця розподілу балів за видами робіт:

Вид контрольного заходу	Питома вага у заліковому модулі	Максимальна оцінка (в балах)
Поточний контроль на лабораторних заняттях (захист лабораторних робіт 1-7)	30 %	100
Виконання самостійної роботи (опрацювання тем 1-6, онлайн-курси)	10 %	100
Модульні контрольні роботи (МКР1 та МКР2 у формі комп'ютерного тестування)	20 % (по 10 % на МКР)	100
Виконання та захист комплексного індивідуального завдання (КПЗ)	10 %	100
Підсумковий контроль (семестровий залік - тестування, теоретичні запитання)	30 %	100

Критерії оцінювання лабораторних робіт та поточного контролю:

90-100 балів – студент виявляє глибокі знання теоретичного матеріалу, володіє практичними навичками побудови SQL-запитів та написання коду на Python, надає вичерпні відповіді на запитання, код працює без помилок;

75-89 балів – студент добре розуміє матеріал, код програми працює відповідно до завдання, але присутні незначні помилки у відповідях або неточності в оптимізації коду;

65-74 балів – студент володіє основними поняттями, проте вагається при відповідях на теоретичні запитання, код програми працює з помірними недоліками або потребує значного доопрацювання;

60-64 балів – студент засвоїв матеріал на мінімально задовільному рівні, код працює з критичними зауваженнями, відповіді на питання є неточними;

35-59 балів – студент погано володіє практичними навичками, код програми містить суттєві логічні помилки і не функціонує належним чином (незадовільно з можливістю повторного складання);

0-34 балів – студент не опрацював матеріал, робота не виконана або код не працює взагалі (незадовільно з обов'язковим повторним вивченням).

Шкала відповідності оцінок ЛНТУ, національних оцінок та оцінок ECTS:

За шкалою ЛНТУ (в балах)	За національною шкалою	За шкалою ECTS
90-100	відмінно	A (відмінно)
85-89	добре	B (дуже добре)
75-84	добре	C (добре)
65-74	задовільно	D (задовільно)
60-64	задовільно	E (достатньо)
35-59 / 0-34	незадовільно	FX / F (незадовільно)

ОРІЄНТОВНІ ТЕСТОВІ ЗАВДАННЯ

1. Яка з перелічених бібліотек є основною для роботи з табличними даними в Python?
 - A. NumPy
 - B. pandas
 - C. Matplotlib
 - D. Scikit-learn
 - E. requests
2. Який метод pandas використовується для зчитування даних із файлу CSV?
 - A. pandas.read_file()
 - B. pandas.open_csv()
 - C. pandas.read_csv()
 - D. pandas.load_csv()
 - E. pandas.import_csv()
3. Яка мова запитів є стандартом для роботи з реляційними базами даних?
 - A. Python
 - B. R
 - C. HTML
 - D. SQL
 - E. NoSQL
4. За допомогою якого SQL-оператора здійснюється об'єднання таблиць на основі спільних значень ключів?
 - A. GROUP BY
 - B. ORDER BY
 - C. MERGE
 - D. JOIN
 - E. UNION
5. Який оператор SQL використовується для фільтрації результатів агрегації після GROUP BY?
 - A. WHERE
 - B. HAVING
 - C. LIMIT
 - D. SELECT
 - E. SORT
6. Яка з перелічених моделей належить до методів навчання з учителем (Supervised Learning)?
 - A. K-Means
 - B. PCA (метод головних компонент)
 - C. Лінійна регресія
 - D. Апріорний алгоритм

E. DBSCAN

7. Яка метрика використовується для оцінки точності моделей класифікації?

A. Mean Squared Error (MSE)

B. R-squared (R²)

C. Accuracy

D. Mean Absolute Error (MAE)

E. Adjusted R²

8. Опишіть призначення великих мовних моделей (LLM), таких як ChatGPT, у Data Science:

A. Тільки побудова дашбордів

B. Автоматизація розрахунків лінійної регресії

C. Генерація аналітичного коду, роз'яснення помилок та структурування звітів

D. Збереження терабайтів сирих даних

E. Візуалізація даних у Power BI

9. Який тип діаграми найкраще підходить для аналізу розподілу однієї числової змінної?

A. Гістограма

B. Кругова діаграма

C. Діаграма розсіювання

D. Стовпчаста діаграма

E. Лінійний графік

10. Що таке «ящик з вусами» (boxplot) у розвідувальному аналізі даних (EDA)?

A. Графік для показу лінійного тренду

B. Візуалізація зв'язку між двома категоріальними ознаками

C. Спосіб графічного представлення розподілу числової змінної через квартилі та виявлення викидів

D. 3D модель даних

E. Таблиця спряженості ознак

11. Яка функція pandas використовується для швидкого отримання основних статистичних характеристик DataFrame (середнє, медіана, мінімум тощо)?

A. df.info()

B. df.describe()

C. df.head()

D. df.shape

E. df.statistics()

12. Для чого використовується оператор loc у pandas?

A. Для локального збереження файлу

B. Для вибірки даних за цілочисловими індексами

C. Для вибірки даних за мітками рядків та назвами стовпців

D. Для підключення до SQL бази

Е. Для видалення стовпців

13. Який метод бібліотеки Scikit-learn використовується для навчання моделі на тренувальних даних?

A. `model.predict()`

B. `model.score()`

C. `model.fit()`

D. `model.train()`

E. `model.evaluate()`

14. Що означає високе значення коефіцієнта детермінації R-squared (наприклад, 0.95)?

A. Модель погано описує дані

B. Модель класифікує з помилкою 95%

C. Модель описує 95% дисперсії цільової змінної

D. Дані містять 95% викидів

E. Модель перенавчена

15. Як обробити пропущені значення в pandas DataFrame, замінивши їх медіаною?

A. `df.dropna()`

B. `df.fillna(df.median())`

C. `df.replace_nan()`

D. `df.astype('median')`

E. `df.drop_cols()`

16. Яка з перелічених бібліотек Python призначена для створення інтерактивних веб-дашбордів?

A. Seaborn

B. Dash

C. Requests

D. NumPy

E. Matplotlib

17. В чому полягає ефект перенавчання (overfitting) моделі машинного навчання?

A. Модель показує однакову якість на тренувальній та тестовій вибірках

B. Модель дуже добре описує тренувальні дані, але показує низьку точність на нових даних

C. Модель взагалі не в змозі знайти закономірності в даних

D. Навчання моделі триває занадто довго

E. Модель вимагає занадто багато пам'яті

18. Яка роль сторітелінгу в аналізі даних?

A. Тільки кодування алгоритмів

B. Створення складних математичних доведень

C. Перетворення аналітичних результатів на зрозумілу історію для прийняття бізнес-рішень

- D. Збереження резервних копій коду
 - E. Написання технічної документації API
19. Опишіть операцію LEFT JOIN в SQL:
- A. Повертає тільки ті рядки, які мають збіги в обох таблицях
 - B. Повертає всі рядки з лівої таблиці та відповідні рядки з правої (з NULL у разі відсутності збігу)
 - C. Повертає всі рядки з обох таблиць незалежно від збігів
 - D. Повертає тільки унікальні значення стовпців
 - E. Видаляє з лівої таблиці дублікати
20. Який інструмент є стандартом де-факто для інтерактивного дослідження даних та прототипування коду в Python?
- A. Notepad++
 - B. Jupyter Notebook
 - C. MS Word
 - D. Command Prompt
 - E. PyCharm Professional
21. Яка метрика є найкращою для оцінки якості класифікації в умовах сильного дисбалансу класів?
- A. Accuracy
 - B. MSE
 - C. F1-score
 - D. R-squared
 - E. MAE
22. Яка основна мета попередньої обробки даних (Data Preprocessing) у Data Science?
- A. Візуалізація фінальних звітів
 - B. Підготовка та очищення даних від шумів, пропусків та викидів для покращення якості моделювання
 - C. Публікація результатів у мережі Інтернет
 - D. Створення бекапів баз даних
 - E. Тільки перейменування файлів
23. Яка бібліотека Python є обгорткою над Matplotlib та надає високорівневий інтерфейс для красивих статистичних графіків?
- A. NumPy
 - B. pandas
 - C. Seaborn
 - D. Scikit-learn
 - E. Dash
24. Як імпортувати таблицю з бази даних SQLite directement у pandas DataFrame?
- A. Використати метод `pandas.read_sql_query()` з підключенням до бази
 - B. Спочатку експортувати в TXT, потім зчитати

С. Використати метод `pandas.import_sql()`

D. Напрямую скопіювати файл бази

E. Це неможливо зробити засобами pandas

25. Який метод машинного навчання використовується для розбиття об'єктів на групи на основі схожості ознак без наявності вчителя (розмітки)?

A. Лінійна регресія

B. Логістична регресія

C. Дерево рішень

D. Кластеризація (наприклад, K-Means)

E. Оцінка F1-score

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Литвин В. В., Пасічник В. В., Нікольський Ю. В. Аналіз даних та знань: навчальний посібник. Львів: Магнолія, 2021. 276 с.
2. Іванчук Я.В., Месюра В.І., Яровий А.А., Манжілевський О.Д. Інтелектуальний аналіз даних та машинне навчання. Частина 1. Базові методи та засоби аналізу даних. Вінниця: ВНТУ, 2021. 69 с.
3. Лупан І.В. Інтелектуальний аналіз даних Data Mining: навчально-методичний посібник. Кропивницький, ФОП Піскова М. А., 2022. 112 с.
4. Wes McKinney. Python for Data Analysis. 3rd Edition. O'Reilly Media, 2022. URL: <https://wesmckinney.com/book>.
5. Jake VanderPlas. Python Data Science Handbook. 2nd Edition. O'Reilly Media, 2022. URL: <https://jakevdp.github.io/PythonDataScienceHandbook/>.
6. Aurélien Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd Edition. O'Reilly Media, 2022. 850 p.
7. Alan Dennis, Barbara Wixom, Roberta M. Roth. Systems Analysis and Design. 8th Edition. Wiley, 2021. 512 p.
8. Prometheus "Аналіз даних та статистичне виведення на мові R" Online Course. URL: <https://prometheus.org.ua>.
9. Coursera "Data Science Fundamentals with Python and SQL Specialization" by IBM. URL: <https://www.coursera.org>.
10. Ліщина В. О., Ліщина Н. М. Методи та прийоми аналізу даних при розв'язанні прикладних ІТ-задач. Тези доповідей Міжнародної конференції. Луцьк: ЛНТУ, 2023. С. 112-115.

11. Самчук Л., Повстяна Ю., Ліщина Н., Кліменко А. Використання UML та аналітичних моделей даних для оптимізації процесів виробництва. Вісник ЛНТУ, 2023. №4. С. 182-189.

Методичне видання

DATA SCIENCE

Методичні вказівки до виконання самостійної роботи для здобувачів першого (бакалаврського) рівня вищої освіти освітньої програми «Комп'ютерні науки» галузі знань F Інформаційні технології спеціальності F3 Комп'ютерні науки денної та заочної форм навчання

Укладач: Валерій Олександрович Ліщина

Комп'ютерний набір та верстка: В. О. Ліщина

Редактор: В. О. Ліщина

Підп. до друку "___" _____ 2025 р. Папір офс.

Гарнітура Таймс. Ум. друк. арк. 1.5. Обл.-вид. арк. 1.6.

Тираж 50 прим.

Відділ іміджу та промоції

Луцького національного технічного університету

43018 м. Луцьк, вул. Львівська, 75

Друк - ВІП ЛНТУ