

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЗ „ЛУГАНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА”



Навчально-науковий інститут математики та інформаційних технологій
Кафедра математики та інформатики

Ю. Г. Козуб, О. О. Смагіна, В. Ю. Козуб

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

**методичні вказівки до лабораторних робіт
для здобувачів вищої освіти першого (бакалаврського) рівня
спеціальності F3 – «Комп'ютерні науки»**

Лубни - 2026

УДК 004.272:37

Рекомендовано до друку вченою радою
ДЗ «Луганський національний університет імені Тараса Шевченка»,
(протокол № 7 від 30.01.2026 р.)

Укладачі:

Козуб Ю. Г. – доктор технічних наук, професор;

Смагіна О. О. – кандидат педагогічних наук доцент;

Козуб В. Ю. – доктор філософії з комп'ютерних наук.

Рецензенти:

Гоменюк С. І. – доктор технічних наук, професор, декан математичного факультету
Запорізького національного університету;

Могильний Г. А. – кандидат технічних наук, доцент, директор навчально-наукового
інституту математики та інформаційних технологій ДЗ „Луганський національний
університет імені Тараса Шевченка”.

Козуб Ю. Г., Смагіна О. О., Козуб В. Ю. Інтелектуальний аналіз даних: методичні
вказівки до лабораторних для здобувачів вищої освіти першого бакалаврського рівня
спеціальності F3 – «Комп'ютерні науки». Лубни: ДЗ „ЛНУ імені Тараса Шевченка”, 2026.
37 с.

Методичні вказівки призначені для здобувачів вищої освіти першого
(бакалаврського) рівня, які навчаються за денною та заочною формами з метою розвитку та
закріплення практичних навичок під час підготовки та виконання лабораторних робіт.
Представлені основні відомості, практичні завдання та контрольні питання до
лабораторних робіт з освітнього компоненту «Інтелектуальний аналіз даних».

© Козуб Ю.Г., Смагіна О.О., Козуб В.Ю., 2026

© ДЗ „ЛНУ імені Тараса Шевченка”, 2026

© Козуб Ю.Г., Смагіна О.О., Козуб В.Ю., 2026

© ДЗ „ЛНУ імені Тараса Шевченка”, 2026

ПЕРЕДМОВА

Мета вивчення дисципліни: Основною метою дисципліни «Інтелектуальний аналіз даних» є формування у здобувачів сучасних уявлень про методи та технології пошуку у великих наборах даних, оволодіння практичними навичками використання алгоритмів машинного навчання та статистичного аналізу при вирішенні реальних аналітичних задач. Також до цілей освоєння дисципліни відносяться формування у здобувачів здібностей з реалізації повного циклу Data Mining проєктів від формулювання бізнес-задачі до впровадження аналітичних рішень у виробничі системи.

В результаті освоєння курсу здобувач опанує: основи процесу Knowledge Discovery in Databases (KDD); методи попередньої обробки, очищення та трансформації даних; алгоритми класифікації, кластеризації та регресії; техніки асоціативного аналізу та виявлення аномалій; методи аналізу часових рядів та прогнозування; технології обробки текстових даних та природної мови; архітектури систем обробки великих даних (Hadoop, Spark); принципи оцінки якості моделей та інтерпретації результатів; стандарти етичного використання даних та забезпечення конфіденційності; сучасні інструменти та бібліотеки для Data Mining (Python, R, scikit-learn, TensorFlow).

Вміти: формулювати аналітичні задачі на основі бізнес-потреб; виконувати дослідницький аналіз даних та візуалізацію; здійснювати очищення даних та обробку пропущених значень; створювати нові ознаки та виконувати feature engineering; обирати оптимальні алгоритми для конкретних типів задач; налаштовувати гіперпараметри моделей машинного навчання; оцінювати якість моделей за допомогою метрик та валідації; інтерпретувати результати та комунікувати, надати висновки стейкхолдерам; застосовувати методи боротьби з дисбалансом класів; впроваджувати моделі у виробництво з використанням MLOps практик.

Володіти: навичками роботи з бібліотеками Python для аналізу даних (pandas, NumPy); методиками побудови предиктивних моделей з

використанням scikit-learn; техніками візуалізації даних за допомогою matplotlib, seaborn, Plotly; інструментами роботи з Big Data по типу (PySpark, Dask); практиками версіонування моделей та експериментів (MLflow, DVC); методами виявлення та усунення bias у моделях; техніками explainable AI для інтерпретації рішень моделей; навичками створення аналітичних звітів та презентацій; інструментами автоматизованого машинного навчання (AutoML); навичками роботи з SQL для витягу та агрегації даних.

ЗМІСТ

ТЕМА 3. МЕТОДИ КЛАСИФІКАЦІЇ ТА РЕГРЕСІЇ: ВІД КЛАСИКИ ДО СУЧАСНОСТІ	11
ТЕМА 4. КЛАСТЕРНИЙ АНАЛІЗ ТА МЕТОДИ НЕНАГЛЯДНОГО НАВЧАННЯ	14
ТЕМА 5. АНАЛІЗ АСОЦІАТИВНИХ ПРАВИЛ ТА РЕКОМЕНДАЦІЙНІ СИСТЕМИ.....	16
ТЕМА 7. ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ТЕКСТОВИХ ДАНИХ	21
ТЕМА 8. DATA MINING У КОНТЕКСТІ BIG DATA: ТЕХНОЛОГІЇ ТА ІНСТРУМЕНТИ ЧАСТИНА 1.....	24
ТЕМА 8. DATA MINING У КОНТЕКСТІ BIG DATA: ТЕХНОЛОГІЇ ТА ІНСТРУМЕНТИ ЧАСТИНА 2.....	26
ТЕМА 9. ЕТИКА, КОНФІДЕНЦІЙНІСТЬ ТА DATA GOVERNANCE У DATA MINING ЧАСТИНА 1.....	28
ТЕМА 9. ЕТИКА, КОНФІДЕНЦІЙНІСТЬ ТА DATA GOVERNANCE У DATA MINING ЧАСТИНА 2.....	30
ТЕМА 10. ПРИКЛАДНІ КЕЙСИ ТА ІНСТРУМЕНТИ DATA MINING ЧАСТИНА 1	32

ТЕМА 1. ВСТУП ДО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ: КОНЦЕПЦІЇ ТА ЕВОЛЮЦІЯ

Мета роботи: набути практичних навичок реалізації повного циклу Knowledge Discovery in Databases (KDD) на реальному датасеті, опанувати методики попередньої обробки та трансформації даних, навчитися застосовувати алгоритми класифікації для розв'язання бізнес-задач, оволодіти технікою оцінки якості моделей та інтерпретації результатів за допомогою професійних інструментів Python.

Практичне завдання

1. Необхідно провести аналіз та розробити модель кредитного скорингу для банківської установи, що надала датасет з 100 клієнтів. Задача полягає у прогнозуванні ймовірності дефолту (прострочки кредиту) на основі характеристик клієнтів. Датасет містить наступні ознаки: вік клієнта (age), річний дохід (income), стаж роботи (employment_years), сума кредиту (loan_amount), кредитна історія (credit_history: good/poor/none), рівень освіти (education), сімейний стан (marital_status), кількість утриманців (num_dependents), наявність власності (has_property), кількість попередніх прострочок (previous_defaults). Цільова змінна – бінарна змінна default (0 = немає прострочки, 1 = є прострочка).

2. Процес включає виконання всіх етапів KDD (Knowledge Discovery in Databases): спочатку необхідно завантажити датасет та провести первинний огляд структури даних, перевірити наявність пропущених значень, дублікатів та аномалій. На етапі очищення потрібно застосувати методи імпутації для заповнення пропусків (медіана для числових змінних, мода для категоріальних), видалити дублікати записів та виявити викиди за допомогою методу міжквартильного розмаху (IQR). Далі виконується дослідницький аналіз даних (EDA) з побудовою візуалізацій: розподіл цільової змінної, кореляційна матриця для числових ознак, box plots для виявлення зв'язків між features та target.

3. Етап feature engineering передбачає створення нових ознак, зокрема debt_to_income_ratio (відношення суми кредиту до доходу), age_group (групування віку на категорії: молодий, середній, старший вік), та інших похідних показників, що можуть підвищити предиктивну силу моделі. Категоріальні змінні необхідно закодувати за допомогою Label Encoding або One-Hot Encoding залежно від їх кардинальності. Числові змінні слід стандартизувати з використанням StandardScaler для приведення до нульового середнього та одиничної дисперсії.

Контрольні питання

1. Опишіть розуміння процесу Knowledge Discovery in Databases (KDD) та його основних етапів, зокрема необхідно пояснити чому Data Mining є лише одним з етапів KDD, а не повним процесом, розкрити різницю між етапами попередньої обробки (preprocessing) та трансформації (transformation) даних, навести конкретний приклад з банківської сфери або e-commerce, де кожен етап KDD відіграє критичну роль, а також пояснити що станеться, якщо пропустити етап оцінки патернів (evaluation) перед впровадженням моделі у виробництво.

2. Поясніть різницю між основними типами задач Data Mining, де необхідно розкрити відмінності між класифікацією та кластеризацією з точки зору наявності міток класів, пояснити коли застосовується регресія замість класифікації і навести приклади, описати як асоціативний аналіз відрізняється від інших типів задач та навести real-world застосування з retail індустрії, а також обґрунтувати чому виявлення аномалій вважається складнішим завданням порівняно з класифікацією та які підходи використовуються для боротьби з дисбалансом класів у цій задачі.

3. Пояснити що таке Real-time Analytics та Edge Computing, чим вони відрізняються від традиційного batch processing, навести приклади індустрій де критично важлива обробка даних у реальному часі, розкрити сутність Explainable AI (XAI) та навести ситуації коли інтерпретованість моделі важливіша за її точність.

Література

1. Zhang Chao, Li Wentao, Zhang Huiyan, Zhan Tao. Recent Advances in Intelligent Data Analysis and Its Applications, 2nd Edition. 2025. URL: <https://www.mdpi.com/2079-9292/14/2/228>
2. De Bie Tijl, De Raedt Luc, Hernández-Orallo José, Hoos Holger H., Smyth Padhraic, Williams Christopher K. I. Automating Data Science: Prospects and Challenges. 2022. URL: <https://arxiv.org/pdf/2105.05699>

ТЕМА 2. ІНТЕГРАЦІЯ МАШИННОГО НАВЧАННЯ ТА ГЛИБОКОГО НАВЧАННЯ В DATA MINING

Мета роботи: набути практичних навичок застосування різних алгоритмів машинного навчання для розв'язання задачі класифікації, освоїти методики порівняльного аналізу моделей за множинними метриками якості, навчитися виконувати гіперпараметричне налаштування та cross-validation, оволодіти технікою інтерпретації результатів та вибору оптимальної моделі відповідно до бізнес-критеріїв.

Практичне завдання

1. Необхідно розробити систему прогнозування відтоку клієнтів (customer churn prediction) для телекомунікаційної компанії. Датасет містить інформацію про 100 клієнтів з наступними характеристиками: customerID (унікальний ідентифікатор), gender (стать), SeniorCitizen (чи є пенсіонером 0/1), Partner (чи має партнера), Dependents (чи має утриманців), tenure (кількість місяців обслуговування), PhoneService (наявність телефонного сервісу), MultipleLines (декілька ліній), InternetService (тип інтернету: DSL/Fiber optic/No), OnlineSecurity, OnlineBackup, DeviceProtection.

2. Процес включає виконання повного циклу ML-проєкту: спочатку завантажити датасет Telco Customer Churn з Kaggle або використати синтетичний датасет, провести EDA з аналізом розподілу churn rate (зазвичай 20-30% дисбаланс класів), побудувати візуалізації залежності відтоку від типу контракту, tenure, MonthlyCharges. Виконати очищення даних, перевірити TotalCharges на коректність типу даних (можливо string замість float через пробіли), обробити пропуски.

3. Навчити та порівняти п'ять різних алгоритмів класифікації: Logistic Regression (baseline модель), Decision Tree Classifier (для інтерпретованості), Random Forest Classifier (ансамбль дерев).

Контрольні запитання

1. Опишіть розуміння різниці між наглядним та ненаглядним навчанням, зокрема необхідно пояснити чому класифікація відноситься до supervised learning, а кластеризація до unsupervised, навести по три реальних приклади бізнес-задач для кожного типу навчання, розкрити які виклики виникають при роботі з unlabeled даними порівняно з labeled, пояснити концепцію semi-supervised learning та навести сценарії коли розмітка лише частини даних може бути ефективною стратегією, а також описати як reinforcement learning відрізняється від supervised/unsupervised та навести приклад його застосування в Data Mining.

2. Поясніть метрики оцінки якості класифікаційних моделей та їх вибір залежно від контексту задачі, де необхідно розкрити що таке True Positives, False Positives, True Negatives, False Negatives на прикладі медичної діагностики, пояснити чому accuracy не є оптимальною метрикою для дисбалансованих даних та навести приклад де 95% accuracy може бути поганим результатом,

3. Пояснити концепцію Transfer Learning та навести приклад як попередньо навчена модель BERT може бути адаптована для sentiment analysis відгуків клієнтів вашого бізнесу з мінімальними даними, розкрити що таке Foundation Models (LLM) та як вони змінюють підходи до Text Mining порівняно з традиційними методами (TF-IDF, Word2Vec).

Література

1. Lin Yuanguo, Chen Hong, Xia Wei, Lin Fan, Wang Zongyue, Liu Yong. A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. 2024. URL: <https://arxiv.org/abs/2309.04761>
2. Wang Senzhang, Cao Jiannong, Yu Philip S. Deep Learning for Spatio-Temporal Data Mining: A Survey. 2019. URL: <https://arxiv.org/abs/1906.04928>

ТЕМА 3. МЕТОДИ КЛАСИФІКАЦІЇ ТА РЕГРЕСІЇ: ВІД КЛАСИКИ ДО СУЧАСНОСТІ

Мета роботи: набути практичних навичок застосування класичних та ансамблевих методів класифікації для розв'язання реальної бізнес-задачі, овоїти методики аналізу важливості ознак та інтерпретації рішень моделей, навчитися обирати оптимальну модель на основі множинних критеріїв включаючи точність, швидкість роботи та інтерпретованість, оволодіти технікою роботи з дисбалансованими даними та методами покращення якості прогнозування.

Практичне завдання

1. Необхідно розробити інтелектуальну систему діагностики серцево-судинних захворювань на основі датасету Heart Disease UCI, що містить медичні показники 10-20 пацієнтів. Датасет включає наступні ознаки: age (вік пацієнта), sex (стать: 1=чоловік, 0=жінка), cp (тип болю в грудях: 0-3), trestbps (артеріальний тиск у спокої мм рт.ст.), exang (стенокардія при навантаженні: 1=так, 0=ні), oldpeak (депресія ST сегменту), slope (нахил піку навантаження ST: 0-2), ca (кількість основних судин 0-3), thal (таласемія: 0=normal, 1=fixed defect, 2=reversible defect). Цільова змінна – target (0=немає хвороби, 1-4=ступінь захворювання), яку потрібно перетворити у бінарну (0=здоровий, 1=хворий).

2. Навчити та порівняти шість алгоритмів класифікації з різною філософією: Decision Tree (для інтерпретованості лікарями, максимальна глибина обмежена до 5 для простоти), k-Nearest Neighbors з k=5 та Euclidean distance, Naive Bayes (Gaussian для неперервних медичних показників), Support Vector Machine з RBF kernel, Random Forest з 5 деревами та class_weight='balanced', Gradient Boosting (XGBoost) з early_stopping для оптимальної кількості ітерацій. Для кожної моделі виконати: розділення даних на train/test у пропорції 75% / 25% з stratify для збереження пропорції захворювань; навчання на тренувальній вибірці з 5-fold Stratified Cross-Validation для оцінки стабільності метрик; обчислення confusion matrix з

аналізом Type I Error (False Positive – здорова людина діагностована як хвора, зайві обстеження) та Type II Error (False Negative – хвора людина не виявлена, критична помилка у медицині); sensitivity критично важливий для виявлення всіх хворих; побудова ROC кривої та обчислення AUC score; аналіз швидкості навчання та передбачення (важливо для real-time систем).

Контрольні запитання

1. Опишіть розуміння алгоритму Decision Trees включаючи критерії розбиття та методи боротьби з overfitting, зокрема необхідно пояснити різницю між Information Gain (Entropy) та Gini Impurity як критеріями вибору атрибуту для розбиття, навести приклад розрахунку обох метрик на конкретному датасеті з 3-4 прикладами, розкрити концепцію pruning (pre-pruning через обмеження глибини та post-pruning через видалення гілок після побудови), пояснити чому дерева рішень є інтерпретованими моделями та коли це критично важливо (медицина, фінанси, юридичні рішення), описати як побудувати ансамбль дерев у Random Forest та чому випадковий вибір підмножини features при кожному розбитті знижує кореляцію між деревами, а також порівняти Random Forest з одним глибоким деревом з точки зору bias-variance tradeoff.

2. Поясніть принцип роботи Support Vector Machine та концепцію kernel trick, де необхідно розкрити що таке hyperplane та margin, чому SVM прагне максимізувати margin між класами, що таке support vectors та чому тільки вони впливають на положення hyperplane, пояснити роль гіперпараметра C (regularization) та навести приклади коли використовувати малі C (simpler model, wide margin) проти великі C (complex model, narrow margin), розкрити сутність kernel trick як способу перетворення нелінійно розділюваних даних у вищій розмірності де вони стають лінійно розділюваними без явного обчислення координат.

Література

1. Diamantopoulou Maria J., Papamichail Dimitris M. Performance Evaluation of Regression-Based Machine Learning Models for Modeling Reference

Evapotranspiration with Temperature Data. 2024. URL:
<https://www.mdpi.com/2306-5338/11/7/89>

2. Trigka Maria, Dritsas Elias. A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection. 2025. URL:
<https://www.mdpi.com/1424-8220/25/1/214>

ТЕМА 4. КЛАСТЕРНИЙ АНАЛІЗ ТА МЕТОДИ НЕНАГЛЯДНОГО НАВЧАННЯ

Мета роботи: набути практичних навичок застосування різних методів кластеризації для сегментації клієнтів, освоїти методики порівняльного аналізу алгоритмів за внутрішніми метриками якості, навчитися виконувати вибір оптимальної кількості кластерів за допомогою статистичних методів, оволодіти технікою візуалізації високовимірних даних через зменшення розмірності та інтерпретації бізнес-характеристик виявлених сегментів.

Практичне завдання

1. Необхідно виконати сегментацію клієнтів для онлайн-ритейлера з метою персоналізації маркетингових стратегій на основі датасету Online Retail Dataset (UCI ML Repository) який містить транзакції інтернет-магазину подарунків за період 2010-2011 років. Датасет включає: InvoiceNo (унікальний номер транзакції, якщо починається з 'C' – то скасування), StockCode (код товару), Description (опис товару), Quantity (кількість, негативна для повернень), InvoiceDate (дата та час транзакції), UnitPrice (ціна за одиницю у фунтах стерлінгів), CustomerID (унікальний ідентифікатор клієнта), Country (країна клієнта). Завдання – кластеризувати клієнтів на основі їх купівельної поведінки для розробки targeted маркетингових кампаній.

2. Провести exploratory data analysis з акцентом на розподіли та кореляції: побудувати histograms для Recency, Frequency, Monetary з логарифмічним масштабом (RFM мають right-skewed розподіли), створити scatter plots для пар (Recency vs Monetary, Frequency vs Monetary) для виявлення візуальних груп, обчислити кореляційну матрицю між усіма features та heat map візуалізацію.

3. Виявити клієнтів-outliers через метод IQR або Z-score (наприклад VIP клієнти з дуже високим Monetary). Виконати трансформацію features: застосувати log transformation для Monetary та Frequency щоб зменшити skewness, стандартизувати всі features за допомогою StandardScaler (критично

для distance-based алгоритмів), опціонально створити polynomial features або interaction terms якщо є підозра на нелінійні залежності.

Контрольні питання

1. Опишіть розуміння алгоритму K-Means включаючи процес збіжності та методи вибору оптимальної кількості кластерів, зокрема необхідно пояснити два основні кроки ітеративного процесу (assignment step та update step) та чому алгоритм гарантовано збігається до локального мінімуму функції WCSS, розкрити проблему локальних мінімумів та як k-means++ ініціалізація допомагає вибирати центроїди що розташовані далеко один від одного для кращого покриття простору, описати Elbow Method з поясненням чому WCSS завжди зменшується зі збільшенням K та як визначити оптимальну точку перегину на графіку.

2. Пояснити Silhouette Score з інтерпретацією значень (s близько до 1 – добре кластеризовано, близько до 0 – на межі кластерів, негативний – можливо неправильний кластер), навести обмеження K-Means щодо форми кластерів (працює тільки зі сферичними кластерами однакового розміру) та навести приклад датасету де K-Means провалиться (концентричні кола, місяці), а також пояснити чому стандартизація features критично важлива для K-Means та що станеться якщо один feature має масштаб 0-1000 а інший 0-1.

3. Поясніть різницю між алгоритмами DBSCAN та K-Means з точки зору їх підходів до кластеризації.

Література

1. Zhou Sheng, Xu Hongjia, Zheng Zhuonan, Chen Jiawei, Li Zhao, Bu Jiajun, Wu Jia, Wang Xin, Zhu Wenwu, Ester Martin. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. 2022. URL: <https://arxiv.org/abs/2206.07579>

2. Ren Yazhou, Pu Jingyu, Yang Zhimeng, Xu Jie, Li Guofeng, Pu Xiaorong, Yu Philip S., He Lifang. Deep Clustering: A Comprehensive Survey. 2022. URL: <https://arxiv.org/abs/2210.04142>

ТЕМА 5. АНАЛІЗ АСОЦІАТИВНИХ ПРАВИЛ ТА РЕКОМЕНДАЦІЙНІ СИСТЕМИ

Мета роботи: набути практичних навичок розроблення та застосування алгоритмів асоціативного аналізу для виявлення patterns у транзакційних даних, освоїти методики Market Basket Analysis та інтерпретації асоціативних правил за метриками support, confidence та lift, навчитися розробляти колаборативну рекомендаційну систему на основі user-item взаємодій, оволодіти технікою оцінки якості рекомендацій та формулювання бізнес-стратегій на основі виявлених асоціацій.

Практичне завдання

1. Необхідно виконати комплексний аналіз покупок у супермаркеті та розробити рекомендаційну систему на основі датасету Groceries Dataset який містить 200 транзакцій з 169 унікальних товарів за місяць роботи продуктового магазину. Кожна транзакція представлена списком куплених items без інформації про ціни або кількості (бінарні дані: куплено/не куплено). Товари згруповані у категорії: dairy products (молочні), vegetables (овочі), fruits (фрукти), bakery (випічка), meat (м'ясо), canned food (консерви), beverages (напої), snacks (снеки), frozen foods (заморожені), household (побутові товари). Завдання включає два компоненти: (1) Market Basket Analysis для виявлення асоціативних правил між товарами, (2) побудова item-based collaborative filtering рекомендаційної системи.
2. Market Basket Analysis через алгоритм Apriori: завантажити та підготувати датасет у форматі транзакційної бази де кожен рядок представляє одну покупку, виконати exploratory analysis для розуміння даних: побудувати розподіл кількості items на транзакцію (середнє, медіана, мода), знайти топ-20 найпопулярніших товарів та побудувати bar chart їх частоти появи.
3. Обчислити sparsity матриці (скільки % можливих комбінацій товарів фактично куплені разом). Застосувати Apriori алгоритм з початковими параметрами $\text{min_support} = 0.01$ (1%, тобто товар має з'явитися принаймні у 98 транзакціях) та $\text{min_confidence} = 0.3$ (30%), виконати sensitivity analysis для

різних threshold: протестувати support від 0.005 до 0.05 та побудувати графік кількості згенерованих правил vs min_support, аналогічно для confidence від 0.2 до 0.8, обрати оптимальні параметри що балансують між кількістю правил (не занадто багато для аналізу) та їх значущістю.

Контрольні питання

1. Необхідно пояснити що показує support та чому низький support (наприклад 0.1%) може бути цінним для luxury товарів але неприйнятним для масових продуктів, розкрити різницю між confidence та conditional probability.
2. Поясніть алгоритм Apriori включаючи Apriori Property та процес генерації кандидатів, де необхідно розкрити принцип антимонотонності support (якщо itemset є infrequent то всі його надмножини також infrequent) та як це використовується для pruning пошукового простору, описати два основні кроки кожної ітерації: join step (генерація candidate k-itemsets з frequent, та prune step (відсіє кандидатів з support нижче threshold), навести покроковий приклад виконання Apriori на малому датасеті (4-5 транзакцій).
3. Потрібно описати підходи до побудови рекомендаційних систем включаючи Collaborative Filtering та Content-Based методи з їх перевагами та недоліками, зокрема пояснити концепцію User-Based Collaborative Filtering як знаходження схожих користувачів та рекомендування items які вони люблять, розкрити формулу predicted rating через weighted average рейтингів схожих користувачів з вагами як similarity scores, описати Item-Based Collaborative Filtering та чому він зазвичай переважає User-Based у production системах (item similarities більш стабільні у часі, можна precompute similarities, краще масштабується), пояснити metrics для обчислення similarity.

Література

1. Sweidan Dirar. Association Rules and Offline-Data-Based Recommender Systems. 2020. URL: <https://www.diva-portal.org/smash/get/diva2:1603361/FULLTEXT01.pdf>
2. Kannout Eyad, Nguyen Hung Son, Grzegorowski Marek. Speeding Up Recommender Systems Using Association Rules. 2022. URL: <https://arxiv.org/abs/2211.08799>

ТЕМА 6. АНАЛІЗ ЧАСОВИХ РЯДІВ ТА ПРОГНОЗУВАННЯ

Мета роботи: набути практичних навичок декомпозиції часових рядів на компоненти (trend, seasonality, residuals), освоїти методики перевірки стаціонарності та трансформації нестационарних рядів, навчитися застосовувати класичні статистичні моделі (ARIMA, SARIMA) та сучасні ML підходи (Prophet, LSTM) для прогнозування, оволодіти технікою оцінки якості прогнозів через time series cross-validation та порівняльного аналізу різних моделей за множинними метриками.

Практичне завдання

1. Необхідно розробити систему прогнозування пасажиропотоку авіакомпанії на основі класичного датасету AirPassengers який містить щомісячні дані про кількість міжнародних авіапасажирів за період з січня 2016 по грудень 2017 року. Датасет демонструє чіткий зростаючий тренд та річну сезонність.

2. Exploratory Data Analysis та декомпозиція: завантажити датасет та створити pandas DataFrame з datetime index, побудувати line plot часового ряду з анотаціями для виявлення візуальних патернів, обчислити базову описову статистику (mean, std, min, max по роках), виконати seasonal subseries plot (box plots для кожного місяця щоб побачити сезонні патерни), виконати декомпозицію через seasonal_decompose з additive та multiplicative моделями, порівняти результати та обрати адекватнішу модель.

3. Тестування стаціонарності та трансформації: виконати Augmented Dickey-Fuller test на оригінальному ряді, інтерпретувати результати (p-value, test statistic, critical values) та зробити висновок про стаціонарність, виконати KPSS test як complementary тест, застосувати log transformation для стабілізації variance: створити log_passengers series та повторити візуалізацію, виконати first-order differencing на log-transformed series для видалення тренду, повторити ADF test на differenced series та підтвердити стаціонарність, виконати seasonal differencing (lag=12) для видалення річної сезонності,

побудувати ACF та PACF plots для differenced series для визначення параметрів ARIMA.

Контрольні питання

1. Опишіть розуміння компонентів часового ряду та методів їх декомпозиції, зокрема необхідно пояснити чотири основні компоненти (trend, seasonality, cyclical, irregular) з конкретними прикладами з різних domain (фінанси, retail, погода), розкрити різницю між сезонністю та циклічними патернами (seasonality має фіксований період) як 12 місяців.

2. Поясніть концепцію стаціонарності часових рядів та методи її досягнення, де необхідно розкрити визначення стаціонарності через три умови (constant mean, constant variance, constant autocovariance), пояснити чому більшість статистичних моделей (ARIMA) припускають стаціонарність та що станеться якщо застосувати їх до нестаціонарних рядів (spurious correlations, unreliable forecasts)

3. Описати ARIMA моделі включаючи процес ідентифікації параметрів та сучасні альтернативи.

Література

1. Hall Timothy, Rasheed Khaled. A Survey of Machine Learning Methods for Time Series Prediction. 2025. URL: <https://www.mdpi.com/2076-3417/15/11/5957>
2. Liu Xinhe, Wang Wenmin. Deep Time Series Forecasting Models: A Comprehensive Survey. 2024. URL: <https://www.mdpi.com/2227-7390/12/10/1504>

ТЕМА 7. ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ТЕКСТОВИХ ДАНИХ

Мета роботи: набути практичних навичок повного циклу обробки текстових даних від preprocessing до modeling, освоїти методики векторизації тексту через TF-IDF та word embeddings, навчитися застосовувати різні алгоритми для sentiment analysis та topic modeling, оволодіти технікою оцінки якості текстових класифікаторів та інтерпретації виявлених тем для бізнес-insights.

Практичне завдання

1. Розробити систему аналізу відгуків клієнтів для платформи електронної комерції на основі датасету Amazon Product Reviews який містить 100 відгуків на різні товари з оцінками від 1 до 5 зірок. Кожен запис включає: reviewerID, productID, reviewText (текст відгуку), rating (1-5), timestamp, helpful (кількість користувачів що відмітили відгук корисним). Завдання складається з трьох компонентів: (1) sentiment classification для автоматичного визначення тональності відгуків, (2) topic modeling для виявлення основних тем в відгуках різних категорій товарів, (3) aspect-based sentiment analysis для детального розуміння що саме клієнти люблять/не люблять у продуктах.

2. Text Preprocessing Pipeline: завантажити датасет та виконати exploratory analysis: побудувати розподіл ratings (bar chart), обчислити середню довжину відгуків по кожному rating level (hypothesis: негативні відгуки зазвичай довші), візуалізувати top-20 найчастіших слів через word cloud окремо для positive (4-5 stars) та negative (1-2 stars) відгуків

3. Sentiment Classification. Перетворити ratings на binary sentiment labels: positive (4-5 stars → 1), negative (1-2 stars → 0), видалити neutral (3 stars) для спрощення binary classification, розділити дані на train (70%), validation (15%), test (15%) з stratification для балансу класів. Виконати feature engineering через TF-IDF vectorization: використати TfidfVectorizer з max_features=5000, ngram_range=(1,2) для unigrams+bigrams, min_df=5 (видалити рідкісні слова), max_df=0.7 (видалити дуже часті), fit на train та transform train/val/test. Навчити та порівняти три класичні ML моделі: (1) Logistic Regression як baseline з

class_weight='balanced', обчислити classification report та confusion matrix; (2) Random Forest з 100 estimators, проаналізувати feature importances (топ-50 найважливіших слів/bigrams для класифікації); (3) XGBoost з early_stopping_rounds=10 на validation set. Для найкращої моделі побудувати ROC curve та обчислити AUC, проаналізувати misclassified examples (які відгуки модель класифікувала неправильно та чому), виконати error analysis: false positives (негативні класифіковані як позитивні) vs false negatives, чи є патерни у помилках (наприклад сарказм, mixed sentiments).

Контрольні питання

1. Опишіть розуміння етапів preprocessing текстових даних та їх вплив на якість моделей, зокрема необхідно пояснити різницю між tokenization на рівні слів vs sub-word (BPE у BERT) vs символів з прикладами коли кожен підхід переважає, розкрити дилему lowercasing через приклади коли регістр важливий (Apple компанія vs apple фрукт, US країна vs us займенник) та коли можна безпечно lowercase, описати різницю між stemming (Porter, Lancaster) та lemmatization (spaCy, WordNet) з прикладами виходу на одному наборі слів ("running", "ran", "better", "mice", "studies"), обґрунтувати чому lemmatization зазвичай переважає stemming для ML моделей (real words, semantically correct) навіть якщо повільніша.

2. Описати задачі sentiment analysis та topic modeling включаючи сучасні підходи та практичне застосування

3. Поясніть методи векторизації тексту від класичних до сучасних включаючи їх переваги та обмеження, де необхідно розкрити Bag of Words як найпростіше представлення з поясненням чому втрачається порядок слів та чому це проблема для деяких задач (наприклад "not bad" vs "bad not")

Література

1. Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, Lifang He. A Survey on Text Classification: From Shallow to Deep Learning. 2021. URL: <https://arxiv.org/abs/2008.00364>

2. Gianni Brauwiers, Flavius Frasincar. A Survey on Aspect-Based Sentiment Classification. 2022. URL: https://arxiv.org/abs/2203.14266?utm_source

ТЕМА 8. DATA MINING У КОНТЕКСТІ BIG DATA: ТЕХНОЛОГІЇ ТА ІНСТРУМЕНТИ ЧАСТИНА 1

Мета роботи: набути практичних навичок роботи з розподіленими обчисленнями для обробки великих даних, освоїти архітектуру Hadoop та HDFS для distributed storage, навчитися використовувати Apache Spark для паралельної обробки даних та machine learning на кластерах, опанувати методи оптимізації performance для Big Data pipelines, розвинути розуміння trade-offs між різними технологіями зберігання та обробки, освоїти практики роботи з PySpark DataFrame API та Spark MLlib.

Практичне завдання

1. Необхідно розробити distributed data processing pipeline для аналізу логів веб-сервера компанії електронної комерції, яка генерує 1 GB логів щодня у форматі Apache Combined Log. Кожен лог містить: IP адресу клієнта, timestamp запиту, HTTP метод та URI, HTTP статус код, розмір відповіді у байтах, referrer URL, user agent string. Треба завантажити дані у HDFS з replication factor три для fault tolerance, створити Spark application для паралельної обробки логів на кластері з десятима workers nodes, виконати ETL pipeline для очищення даних видаляючи bot traffic за user agent patterns та invalid requests з статус кодами поза діапазоном двісті-п'ятсот, парсингу timestamp у datetime формат та витягу features такі як година дня, день тижня, категорія продукту з URI.

2. Виконати exploratory data analysis через Spark DataFrame API обчислюючи статистику трафіку: загальна кількість requests, unique visitors за IP адресами з використанням approxCountDistinct для ефективності на великих даних, розподіл requests по годинах дня та днях тижня для виявлення peak hours, топ-десять найпопулярніших сторінок за кількістю pageviews, топ-десять referrer domains що приводять найбільше трафіку, average response size та total bandwidth споживання.

3. Розподілити HTTP статус кодів з підрахунком відсотка помилок чотириста-х та п'ятсот-х. Результати агрегацій зберегти у Parquet формат для

ефективного зберігання та подальших запитів, створити partitioned structure по даті для оптимізації queries що фільтрують за датою.

Контрольні питання

1. Опишіть архітектуру HDFS включаючи ролі NameNode та DataNodes з поясненням як зберігаються великі файли через розбиття на blocks розміром сто двадцять вісім мегабайт, розкрити механізм replication для fault tolerance та як HDFS забезпечує data locality principle переміщуючи computation до даних а не навпаки, пояснити read та write process у HDFS з деталями як client взаємодіє з NameNode для metadata та з DataNodes для actual data.

2. Описати rack awareness strategy для розміщення replicas на різних racks для захисту від rack failures, розглянути limitations HDFS такі як high latency not suitable для real-time, inefficient для багатьох малих файлів, NameNode як single point of failure та як High Availability setup з Standby NameNode вирішує цю проблему.

3. Поясніть різницю між MapReduce та Apache Spark з точки зору execution model.

Література

1. Gandomi A. H., Chen Fang, Laith. A. Machine Learning Technologies for Big Data Analytics. 2022. URL: <https://www.mdpi.com/2079-9292/11/3/421>

2. Inoubli Wissem, Aridhi Sabeur, Mezni Haithem, Maddouri Mondher, Mephu Nguifo Engelbert. An Experimental Survey on Big Data Frameworks. 2016. URL: <https://arxiv.org/abs/1610.09962>

ТЕМА 8. DATA MINING У КОНТЕКСТІ BIG DATA: ТЕХНОЛОГІЇ ТА ІНСТРУМЕНТИ ЧАСТИНА 2

Мета роботи: набути практичних навичок роботи з NoSQL базами даних для різних типів Big Data workloads, освоїти streaming data processing для real-time analytics, навчитися проєктувати data pipelines з використанням Apache Kafka для event streaming, опанувати методи оптимізації storage через columnar formats такі як Parquet та ORC, розвинути розуміння Lambda та Kappa architectures для поєднання batch та streaming processing, оволодіти cloud-based Big Data platforms.

Практичне завдання

1. Необхідно спроектувати та реалізувати real-time analytics system для моніторингу транзакцій fraud detection у банківській системі яка обробляє 10 транзакцій за хвилину. Система повинна включати Apache Kafka як distributed streaming platform з трьома brokers для стабільності системи та, producer applications що публікують транзакції, у Kafka topic з partitioning за customer ID для збереження порядку транзакцій одного клієнта.

2. Consumer applications на базі Spark Structured Streaming що обробляють транзакції у мікро патчах (обмін повідомленнями між серверами) з trigger interval у одну хвилину. Event schema містить: transaction ID, customer ID, merchant ID, amount, currency, timestamp, location coordinates, device fingerprint, IP address. Налаштувати Kafka з replication factor, compression type snappy для ефективного використання disk space.

3. Розробити Spark Structured Streaming application яка виконує windowed aggregations для виявлення suspicious patterns. Для кожного виявленого suspicious event публікувати alert у окремий Kafka topic alerts для real-time notification system, зберігати aggregated metrics у txt. форматі.

Контрольні питання

1. Опишіть NoSQL database categories та їх use cases включаючи key-value stores такі як Redis для caching та session storage з поясненням їх extreme

performance через in-memory storage, document databases такі як MongoDB для flexible schema та nested documents ideal для content management systems.

2. Що таке Cassandra, та опишіть як її використовують у даному контексті.

3. Поясніть Apache Kafka architecture та його роль у event-driven architectures де Kafka діє як distributed commit log, розкрити концепції «topics partitions replicas»

Література

1. Gandomi Amir H., Chen Fang, Abualigah Laith. Machine Learning Technologies for Big Data Analytics. 2022. URL: <https://www.mdpi.com/2079-9292/11/3/421>

2. Inoubli Wissem, Aridhi Sabeur, Mezni Haithem, Maddouri Mondher, Mephu Nguifo Engelbert. An Experimental Survey on Big Data Frameworks. 2016. URL: <https://arxiv.org/abs/1610.09962>

ТЕМА 9. ЕТИКА, КОНФІДЕНЦІЙНІСТЬ ТА DATA GOVERNANCE У DATA MINING ЧАСТИНА 1

Мета роботи: вивчити етичні аспекти інтелектуального аналізу даних, принципи забезпечення конфіденційності персональних даних, методи знеособлення інформації та вимоги законодавства щодо захисту даних. Навчитися виявляти та усувати упередженість в моделях машинного навчання, а також застосовувати техніки диференційної конфіденційності для захисту чутливої інформації.

Практичне завдання

1. Компанія з медичного страхування збирається використовувати алгоритми аналізу даних для оцінки ризиків та визначення страхових тарифів на основі медичних даних клієнтів. Необхідно розробити етичну стратегію використання даних та забезпечити відповідність законодавству.

2. Розробити модель прогнозування страхових ризиків на знеособлених даних з використанням алгоритмів регресії або дерев рішень

Встановити обмеження на використання захищених атрибутів: заборонити пряме використання віку, статі, етнічної приналежності у фінальних рішеннях. Застосувати техніки справедливого машинного навчання: рівність можливостей для різних груп, вирівнювання показників точності між демографічними категоріями, мінімізація різниці у помилках для захищених груп.

3. Обчислити метрики справедливості моделі: демографічний паритет (однакові ставки позитивних прогнозів для всіх груп), рівність шансів (однакові показники істинно позитивних результатів), калібрування за групами (точність прогнозів однакова для всіх категорій)

Контрольні питання

1. Опишіть основні етичні проблеми використання алгоритмів аналізу даних у чутливих сферах (медицина, фінанси, правосуддя). Відповідь має включати: види упередженості у даних та алгоритмах (історична упередженість, упередженість вибірки, упередженість підтвердження); ризики

дискримінації та несправедливих рішень щодо вразливих груп населення; проблеми прозорості та пояснюваності складних моделей (ефект чорної скриньки); конфлікт між точністю моделей та етичними принципами; відповідальність за помилкові рішення алгоритмів; приклади реальних випадків етичних порушень у різних галузях.

2. Поясніть методи знеособлення персональних даних та оцініть їх ефективність для захисту конфіденційності. Відповідь повинна охоплювати: псевдонімізацію (заміна ідентифікаторів на псевдоніми зі збереженням можливості відновлення зв'язку); анонімізацію (незворотне видалення ідентифікаторів).

3. Поясніть що таке «узагальнення та придушення даних» (заміна конкретних значень на діапазони або категорії); додавання статистичного шуму до даних; методи диференційної конфіденційності та їх математичні гарантії; модель відповідності та унікальності для оцінки ризику повторної ідентифікації; обмеження різних методів та можливі атаки на знеособлені дані; баланс між корисністю даних та рівнем захисту конфіденційності.

Література

1. Busari M. Ethics, Privacy, and Governance in Business Data Analytics. 2025. URL: https://www.researchgate.net/publication/393663331_Ethics_Privacy_and_Governance_in_Business_Data_Analytics
2. Mirishli S. Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy. 2025. URL: <https://arxiv.org/abs/2503.14539>

ТЕМА 9. ЕТИКА, КОНФІДЕНЦІЙНІСТЬ ТА DATA GOVERNANCE У DATA MINING ЧАСТИНА 2

Мета роботи: вивчити принципи управління даними в організації, методи забезпечення якості та відповідності нормативним вимогам, стратегії корпоративної відповідальності у використанні технологій аналізу даних. Навчитися розробляти політики управління даними, проводити аудит використання даних та створювати системи контролю за дотриманням етичних стандартів.

Практичне завдання

1. Велика роздрібна мережа впроваджує систему аналізу поведінки покупців для персоналізації пропозицій та прогнозування попиту. Необхідно розробити комплексну стратегію управління даними, яка забезпечить відповідність законодавству, етичним стандартам та корпоративним цінностям компанії.

2. Розробка політики управління даними. Створити класифікацію даних за рівнями чутливості: публічні дані (загальна інформація про товари, знеособлена статистика); внутрішні дані (аналітичні звіти, бізнес-метрики); конфіденційні дані (історія покупок клієнтів, платіжні дані); строго конфіденційні дані (медична інформація, біометричні дані). Визначити правила доступу до даних для різних категорій користувачів: аналітики даних (доступ до знеособлених агрегованих даних); менеджери з маркетингу (доступ до сегментів клієнтів без персональних ідентифікаторів); технічний персонал (обмежений доступ для підтримки систем); керівництво (доступ до узагальнених звітів)

3. Розробити життєвий цикл даних: збір (джерела, методи отримання згоди); зберігання (терміни зберігання, умови безпеки); обробка (дозволені види аналізу, обмеження використання); передача (умови передачі третім сторонам); видалення (процедури безпечного знищення після закінчення терміну зберігання).

Контрольні питання

1. Опишіть основні компоненти системи управління даними в організації. Відповідь має включати: політики та процедури роботи з даними (класифікація, доступ, зберігання); організаційну структуру управління даними (ролі та відповідальність, комітети з управління даними); технічну інфраструктуру (системи контролю доступу, засоби шифрування, резервне копіювання); процеси забезпечення якості даних (валідація, очищення, моніторинг); механізми контролю відповідності (аудити, оцінка ризиків, звітність); між управлінням даними та загальною стратегією компанії; метрики ефективності управління даними.

2. Поясніть концепцію оцінки впливу на конфіденційність та наведіть приклади ситуацій, коли вона обов'язкова. Відповідь повинна охоплювати: визначення оцінки впливу на конфіденційність як систематичного процесу виявлення та мінімізації ризиків для прав людей; основні етапи проведення оцінки (опис обробки, оцінка необхідності, виявлення ризиків, заходи зменшення ризиків); критерії, коли оцінка є обов'язковою (нові технології, масштабна обробка чутливих даних).

Література

1. Busari M. Ethics, Privacy, and Governance in Business Data Analytics. 2025. URL: https://www.researchgate.net/publication/393663331_Ethics_Privacy_and_Governance_in_Business_Data_Analytics
2. Mirishli S. Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy. 2025. URL: <https://arxiv.org/abs/2503.14539>

ТЕМА 10. ПРИКЛАДНІ КЕЙСИ ТА ІНСТРУМЕНТИ DATA MINING

ЧАСТИНА 1

Мета роботи: вивчити реальні приклади застосування методів інтелектуального аналізу даних у різних галузях економіки, освоїти підходи до вирішення бізнес-задач за допомогою аналітичних технологій. Навчитися адаптувати типові рішення до специфіки конкретної предметної області, оцінювати ефективність впровадження аналітичних систем та формулювати рекомендації для бізнесу.

Практичне завдання

1. Підготовка та дослідницький аналіз даних. Зібрати дані з різних джерел: транзакційна система (чеки продажів з детальною інформацією про товари, ціни, знижки, дату та час покупки), програма лояльності (профілі клієнтів, історія покупок, використання бонусів), система управління запасами (залишки товарів, дати постачань, терміни придатності), календар маркетингових активностей (акції, святкові періоди). Виконати первинний огляд даних: обсяг даних (кількість транзакцій, унікальних товарів, клієнтів), часовий діапазон, пропущені значення, викиди у цінах або кількостях

2. Провести дослідницький аналіз з візуалізацією: динаміка продажів у часі (тренди, сезонні коливання, аномалії), розподіл покупок за днями тижня та годинами доби, найпопулярніші категорії та окремі товари, аналіз кошиків покупців (середня кількість позицій, середня сума чека), географічний розподіл продажів за магазинами

3. Виявити особливості даних: святкові піки продажів певних категорій, вплив погодних умов на попит, кореляції між категоріями товарів, сегменти клієнтів з різною поведінкою

Контрольні запитання

1. Опишіть процес трансформації бізнес-проблеми у аналітичне завдання для інтелектуального аналізу даних. Відповідь має включати: методи збору вимог від зацікавлених сторін (інтерв'ю, воркшопи, аналіз документів); виявлення ключових бізнес-показників та цілей оптимізації.

2. Дайте визначення «формулюванню конкретних аналітичних питань», на які потрібно відповісти; вибір відповідних методів аналізу даних (класифікація, регресія, кластеризація тощо); визначення критеріїв успіху проекту у вимірних показниках; оцінку доступності необхідних даних та можливості їх отримання; узгодження очікувань щодо термінів та результатів; приклади типових бізнес-проблем та відповідних аналітичних підходів у різних галузях.

Література

1. Tekdogan Taha, Cakmak Ali. Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification. 2022. URL: <https://arxiv.org/abs/2209.10637>
2. Kamath Chandrika, Franzman Juliette, Ponmalai Ravi. Data Mining for Faster, Interpretable Solutions to Inverse Problems: A Case Study Using Additive Manufacturing. 2023. URL: <https://arxiv.org/abs/2306.04228>

ТЕМА 10. ПРИКЛАДНІ KEYСИ ТА ІНСТРУМЕНТИ DATA MINING ЧАСТИНА 2

Мета роботи: освоїти сучасні інструменти та платформи для інтелектуального аналізу даних, навчитися вибирати відповідні технології для конкретних завдань. Отримати практичні навички роботи з популярними бібліотеками машинного навчання, засобами візуалізації даних та хмарними сервісами аналітики.

Практичне завдання

1. Стартап у сфері фінансових технологій розробляє додаток для автоматичного аналізу особистих фінансів користувачів. Необхідно створити повний аналітичний конвеєр: від завантаження даних до розгортання моделі у продуктивному середовищі.

2. Встановити середовище для роботи з даними: мова програмування Python з пакетним менеджером, інтерактивне середовище Jupyter Notebook або аналогічні інструменти для експериментування, система контролю версій Git для відстеження змін у коді. Встановити основні бібліотеки для аналізу даних: бібліотека для роботи з таблицями та обробки даних (наприклад, pandas), бібліотеки для числових обчислень (numpy), інструменти для візуалізації (matplotlib, seaborn), бібліотеки машинного навчання (scikit-learn для класичних алгоритмів). Налаштувати робочий простір: структура каталогів проекту (папки для даних, програмного коду, блокнотів, результатів), файли конфігурації середовища (список залежностей проекту), документація початкового налаштування. Вибрати платформу для зберігання даних: локальні файли для малих обсягів, реляційна база даних для структурованих даних, хмарне сховище для великих обсягів (Google Cloud Storage). Навести приклад чому вибір пав саме на даний тип інструментів: критерії відбору (доступність, вартість, масштабованість, підтримка спільноти), порівняння альтернативних рішень

3. Завантажити синтетичний набір даних про фінансові транзакції користувачів: дата та час операції, категорія витрат (їжа, транспорт, розваги, комунальні послуги, здоров'я тощо), сума транзакції, тип операції (дебет або кредит), опис транзакції, баланс рахунку після операції

Використати бібліотеку `pandas` для читання даних з різних форматів (CSV, Excel, JSON) та створення табличного представлення.

Контрольні запитання

1. Опишіть основні бібліотеки та інструменти екосистеми Python для аналізу даних та машинного навчання, їх призначення та типові сценарії використання. Відповідь має включати: `pandas` - робота з табличними даними (читання, очищення, перетворення, агрегація); `numpy` - ефективні числові обчислення з масивами та матрицями; `Seaborn` та `Matplotlib` - створення статичних візуалізацій даних; а також препроцесинг даних, оцінка моделей.

2. Поясніть концепцію конвеєру обробки даних та переваги його використання у проектах машинного навчання. Відповідь повинна охоплювати: визначення конвеєру як послідовності кроків перетворення даних та навчання моделі; етапи типового конвеєру (завантаження даних, очищення, створення ознак, нормалізація, навчання моделі, оцінка); переваги конвеєрів: відтворюваність експериментів (один і той самий процес для різних запусків), зниження ризику витоку даних між навчальною та тестовою вибірками, спрощення налаштування параметрів усього процесу одночасно, легке розгортання у продуктивне середовище; реалізація конвеєрів у `scikit-learn` за допомогою класу `Pipeline`; приклад створення конвеєру з препроцесингом та класифікацією; автоматизація конвеєрів для регулярного перенавчання моделей; моніторинг роботи конвеєру у продуктивному середовищі.

Література

1. Tekdogan Taha, Cakmak Ali. Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification. 2022. URL: <https://arxiv.org/abs/2209.10637>

2. Kamath Chandrika, Franzman Juliette, Ponmalai Ravi. Data Mining for Faster, Interpretable Solutions to Inverse Problems: A Case Study Using Additive Manufacturing. 2023. URL: <https://arxiv.org/abs/2306.04228>

Навчальне видання

**Юрій Гордійович Козуб
Ольга Олександрівна Смагіна
Владислав Юрійович Козуб**

**ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ
методичні вказівки до лабораторних робіт
для здобувачів вищої освіти першого (бакалаврського) рівня
спеціальності ФЗ – «Комп'ютерні науки»**

Комп'ютерне макетування – О. О. Смагіна
Коректор – Ю. Г. Козуб

Здано до склад. 23.01.2026 р. Підп. до друку 30.01.2026 р. Формат 60х84 1/16.
Папір офсет. Гарнітура Times New Roman. Друк ризографічний.
Ум. друк. арк. 1,2. Наклад 100 прим. Зам. № 131
Видавництво Державний заклад «Луганський національний університет імені Тараса
Шевченка»
м. Полтава, вул. Ковалю, 3
e-mail: mail@luguniv.edu.ua www/luguniv.edu.ua
Свідоцтво суб'єкта видавничої справи ДК № 3459 від 09.04.2009 р.
