**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ**
**Факультет інформаційних технологій**
**Кафедра цифрової економіки та системного аналізу**

# СИЛАБУС (SYLLABUS)
# Дисципліна «Технології аналізу даних»/ Data analysis technologies»

## ІНФОРМАЦІЯ ПРО ВИКЛАДАЧА

| | |
|---|---|
| Викладач | Роскладка Андрій Анатолійович |
| Науковий ступінь | Доктор економічних наук наук |
| Вчене звання | Професор |
| Посада | Завідувач кафедри цифрової економіки та системного аналізу |
| Адреса кафедри | м.Київ, вул. Кіото 19, каб. Б-517, Б-519 |
| E-mail | desa@knute.edu.ua |
| Консультації | Відповідно до графіку індивідуальних консультацій на сайті кафедри |

## ПОЛІТИКА АКАДЕМІЧНОЇ ДОБРОЧЕСНОСТІ

https://knute.edu.ua/file/NjY4NQ==/bf27ad9293fa2bb6f9b2c3031d4b6e4a.pdf

**Дотримання академічної доброчесності передбачає:**

- самостійне виконання навчальних завдань, завдань поточного та підсумкового контролю результатів навчання (для осіб з особливими освітніми потребами ця вимога застосовується з урахуванням їхніх індивідуальних потреб і можливостей);
- посилання на джерела інформації у разі використання не авторських ідей, розробок, тверджень, відомостей і т.п.;
- дотримання норм законодавства про авторське право і суміжні права;
- надання достовірної інформації про результати власної наукової діяльності, використані методики досліджень і джерела інформації.

**Порушенням академічної доброчесності вважається:**

- академічний плагіат – оприлюднення (частково або повністю) наукових (творчих) результатів, отриманих іншими особами, як результатів власного дослідження (творчості) та/або відтворення опублікованих текстів (оприлюднених творів мистецтва) інших авторів без зазначення авторства;
- самоплагіат – оприлюднення (частково або повністю) власних раніше опублікованих наукових результатів як нових наукових результатів;
- фабрикація – вигадування даних чи фактів, що використовуються в наукових дослідженнях;
- фальсифікація – свідома зміна чи модифікація вже наявних даних, що стосуються наукових досліджень.

**За порушення академічної доброчесності здобувачі освіти можуть бути притягнені до академічної відповідальності:**

- повторне проходження оцінювання (модульний контроль, іспит, залік тощо);
- повторне проходження відповідного освітнього компонента освітньо-професійної програми;
- відрахування з Університету;
- позбавлення наданих університетом пільг;
- відмова у присудженні відповідного ступеня вищої освіти;

## ПОЛІТИКА ЩОДО ВІДВІДУВАННЯ ЗАНЯТЬ

- відвідування занять є обов'язковим;
- Студент, який пропустив практичне заняття, самостійно вивчає матеріал (при виникненні питань може звертатися за консультацією згідно розкладу консультацій викладачів оприлюдненого на сайті кафедри) за наведеними джерелами, виконує завдання і здає його викладачу.
- за об'єктивних причин (наприклад, хвороба, міжнародне стажування та ін.) навчання може відбуватись в он-лайн формі за погодженням із викладачем дисципліни.

## ОПИС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

| | |
|---|---|
| **Назва дисципліни/ тип дисципліни** | Технології аналізу даних / вибіркова |
| **Навчальний рік** | 2023-2024 |
| **Факультет** | Факультет інформаційних технологій |
| **Курс** | 4 |
| **Семестр** | 7-8 |
| **Освітній ступінь** | Бакалавр |
| **Галузь знань** | 12 «Інформаційні технології» |
| **Спеціальність** | 122 «Комп'ютерні науки» |
| **Загальна характеристика** | Кількість годин –180<br>Кількість кредитів – 6<br>**Види занять:** лекції, лабораторні, самостійна робота.<br>**Співвідношення аудиторних годин і годин самостійної роботи** - 68/112<br>**Мова викладання** – українська<br>**Форма викладання** – очна |
| **Підсумковий контроль** | Екзамен |
| **Програмне забезпечення** | Microsoft Access Databases, JSON, SQL, Oracle Databases, ODBC, Google Analytics, Azure SQL Database. |
| **Обладнання** | Проектор, комп'ютерна техніка із встановленим програмним забезпеченням та доступом до мережі Інтернет. |
| **Необхідні попередні дисципліни** | • Основи інформаційних технологій (операційна система Windows, бази даних, системи захисту інформації).<br>• Основ дискретної математики, математичної логіки, алгоритмізації та програмування.<br>• Теорія ймовірності та математична статистика. |
| **Методика вивчення** | Методика вивчення дисципліни полягає у набутті студентами знань теоретичного і практично-прикладного характеру під час лекцій, лабораторних занять, самостійної роботи та вивчення першоджерел і навчально-методичної літератури. |
| **Мета і завлання** | **Метою** вивчення дисципліни «Технології аналізу даних» є надання фундаментальних теоретичних знань і набуття практичних навичок з питань формування, дослідження та всебічного аналізу даних у різних галузях сферах людської діяльності.<br>**Завданням** вивчення дисципліни «Технології аналізу даних» є надання студентам ґрунтовних знань в області аналітичних досліджень інформаційного простору, вивчення методів створення, добування, консолідації, переробки, трансформації та аналізу даних. |
| **Місце дисципліни в освітньо-професійній програмі** | |
| **Загальні компетентності** | ЗК 1 Здатність до абстрактного мислення, аналізу та синтезу<br>ЗК 7 Здатність до пошуку, оброблення та аналізу інформації з різних джерел |

| | |
|---|---|
| **Фахові компетентності (результати навчання)** | СК 1 Здатність до математичного формулювання та досліджування неперервних та дискретних математичних моделей, обґрунтовування вибору методів і підходів для розв'язування теоретичних і прикладних задач у галузі комп'ютерних наук, аналізу та інтерпретування<br>СК 2 Здатність до виявлення статистичних закономірностей недетермінованих явищ, застосування методів обчислювального інтелекту, зокрема статистичної, нейромережевої та нечіткої обробки даних, методів машинного навчання та генетичного програмування тощо<br>СК 7 Здатність застосовувати теоретичні та практичні основи методології та технології моделювання для дослідження характеристик і поведінки складних об'єктів і систем, проводити обчислювальні експерименти з обробкою й аналізом результатів<br>СК 11 Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач. |
| **Програмні результати навчання** | ПР 3 Використовувати знання закономірностей випадкових явищ, їх властивостей та операцій над ними, моделей випадкових процесів та сучасних програмних середовищ для розв'язування задач статистичної обробки даних і побудови прогнозних моделей<br>ПР 4 Використовувати методи обчислювального інтелекту, машинного навчання, нейромережевої та нечіткої обробки даних, генетичного та еволюційного програмування для розв'язання задач розпізнавання, прогнозування, класифікації, ідентифікації об'єктів керування тощо<br>ПР 12 Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використаннямпрограмнихінструментівпідтримкибагатовимірногоаналізу даних на основі технологій DataMining, TextMining, WebMining. |

## ТЕМАТИКА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

### Тема 1. Передобробка даних.

Історія розвитку *Artificial Intelligence* і *Business Intelligence*. Характеристика фахівця з аналізу даних. *Softskills* та *hardskills* аналітика даних. Приклади застосування *DataScience* у різних галузях людської діяльності. Етапи розв'язування задач аналізу даних: висунення гіпотез, збір і систематизація даних, побудова моделі, яка пояснює факти, тестування моделі та інтерпретація результатів, застосування отриманої моделі. Технологія *Knowledge Discovery in Databases*. Формування вибірки даних. Консолідація даних. *ETL*-процес. Очищення даних. Трансформація даних. Технологія *DataMining*. Задачі *DataMining*: класифікація, регресія, кластеризація, асоціація, послідовність даних. Поняття про аналітичні системи. Актуальні бізнес-задачі аналітики даних.

### Тема 2. Асоціація даних.

Афінітивний аналіз. Поняття типової транзакції. Предметний набір. Основні поняття *Rules Mining*. Асоціативні правила. Умова та наслідок асоціативного правила. Підтримка та достовірність правил. Значущість асоціативних правил. Міри корисності правил. Ліфт, левередж та покращення асоціативних правил. Алгоритм *apriori*. Пошук предметних наборів. Генерація асоціативних правил. Ієрархічні асоціативні правила. Методи пошуку ієрархічних асоціативних правил. Послідовні шаблони. Програмні засоби пошуку асоціативних правил. Практичний аспект застосування технології асоціативних правил. Секвенціальний аналіз.

### Тема 3. Кластеризація даних.

Формальна постановка задачі кластеризації. Задачі кластеризації: вивчення даних, полегшення аналізу, стиснення даних, прогнозування, виявлення аномалій. Приклади кластеризації в різних областях знань. Представлення результатів кластеризації даних. Базові алгоритми кластеризації. Алгоритм кластеризації *k-means*. Критерій збіжності алгоритму. Міри відстаней у кластеризації.

Міри Евкліда і Манхеттена. Алгоритм *g-means*. Кластеризація за Гюстафсоном-Кесселем. Програмні засоби кластеризації та сегментації даних. Мережі Кохонена. Самоорганізуючі карти Кохонена. Методика побудови карти Кохонена. Вибір кількості нейронів карти. Алгоритм навчання мережі Кохонена. Ієрархічні алгоритми кластерного аналізу. Міри схожості. Методи об'єднання та зв'язку. Ітеративні алгоритми кластерного аналізу. Адаптивні методи кластеризації. Проблеми алгоритмів кластеризації. Невизначеність у виборі критерія якості кластеризації. Проблеми машинних ресурсів. Задача вибору кількості кластерів.

**Тема 4. Класифікація та регресія даних.**

Огляд методів класифікації. Точність класифікації. Оцінювання класифікаційних методів. Кореляційно-регресійний аналіз. Статистичні методи аналізу. Байесівська класифікація. Лінійна регресія. Регресія з категоріальними вхідними змінними. Методи відбору змінних до регресійної моделі. Обмеження у застосуванні регресійних моделей. Використання фіктивних змінних. Логістична регресія. Оцінки максимальної правдоподібності. Значущість вхідних змінних. Використання логістичної регресії для розв'язування задач класифікації. Тест Чоу. *ROC*-аналіз. Множинна логістична регресія. Простий байесівській класифікатор. Методи прогнозування даних. Часовий ряд та його компоненти. Моделі прогнозування часових рядів. Дерева рішень. Структура дерева рішень. Алгоритми побудови дерев рішень. Міри ефективності дерев рішень. Критерії вибору найкращих атрибутів розгалуження. Регресійне дерево рішень. Спрощення дерев рішень. Перенавчання і складність моделей. Критерії оптимізації дерев рішень. Відсікання гілок. Регуляризаційні мережі.

**Тема 5. Технології інтелектуальної обробки даних**

Візуальний аналіз даних *Visual Mining*. Характеристики засобів візуалізації даних. Методи візуалізації. Методи геометричних перетворень. Методи, орієнтовані на пікселі. Ієрархічні образи. Задача та етапи аналізу текстів *Text Mining*. Методи класифікації текстових документів. Видалення стоп-слів. Стеммінг. *N*-грами. Методи кластеризації текстових документів: ієрархічні, бінарні. Задача анотування текстів. Пошук асоціацій. Первинний витяг ключових понять. Ідея *Data Mining* у реальному часі *Real-Time Mining*. Адаптація системи до загальної концепції. Рекомендаційні машини. Класифікація рекомендаційних машин. Агентне навчання. Проблеми аналізу даних з мережі Інтернет. Етапи *Web Mining*. Категорії *Web Mining*. Аналіз використання веб-ресурсів. Використання веб-структур та веб-контенту. Аналіз структури сегмента мережі. Персоналізація інформації. Пошук шаблонів в поведінці користувачів.

**Тема 6. Інструментальні засоби аналізу даних.**

Програмне забезпечення в області аналізу даних. Аналітичні платформи *Deductor Studio, Loginom, RapidMiner, Tableau, Weka, Orange, NodeXL, Qlik*. Технології аналізу даних у продуктах *Microsoft Corporation*. Технологія моделювання даних у *Microsoft Power Pivot*. Інтерактивний інструмент Microsoft *Power View* для дослідження, графічного відображення та представлення даних. Надбудова *Microsoft Power Query* в задачах бізнес-аналітики. Хмарні технології *Microsoft* для аналізу та візуалізації даних. Організація бізнес-аналітики рівня *Business Intelligence (BI)*. Платформа *Microsoft BI*. Механізм аналітики в пам'яті *xVelocity*. Налаштування *Power BI* середовища. Інтерфейс *Power BI Desktop*. Завантаження даних *Power BI* з різних інформаційних джерел. Імпорт даних із реляційних баз даних, текстового файлу, вхідного каналу даних та сервісів аналізу.

**Тема 7. Створення моделі даних.**

Зміна даних у *Power Query*. Трансформація, очищення та фільтрування даних у *Power BI*. Об'єднання даних. Додавання даних. Розщеплення даних. Приведення даних до необхідної форми. Групування та агрегування даних. Створення зв'язків таблиці. Схеми зірки та сніжинки у *Power BI Desktop*. Денормалізація даних у моделі. Створення зручної моделі. Мова запитів *DAX*. Оператори *DAX*. Робота з текстовими функціями. Використання функцій дати та часу *DAX*. Використання інформаційних та логічних функцій. Отримання даних із суміжних таблиць. Використання математичних, тригонометричних та статистичних функцій у *DAX*. Створення обчислювальних стовпців і мір у *Power BI*. Зміна контексту запиту. Використання функцій фільтра у створених мірах. Аналіз часових даних у моделі даних *Power BI*. Створення таблиці дат. Оцінки на основі часового періоду. Зміна контексту дати. Використання функцій дати і часу. Створення напівадитивних мір.

**Тема 8. Побудова аналітичних звітів.**

Створення таблиць та матриць візуалізації даних у *Power BI Desktop*. Побудова стрічкових, кругових діаграм та гістограм. Побудова лінійних та точкових діаграм. Створення візуалізацій на основі карт. Поєднання візуалізацій у *Power BI*. Деталізація візуалізацій. Публікація звітів та створення інформаційних панелей на порталі *Power BI*. Опублікування файлів *Power BI Desktop* у *Power BI Service*. Додавання плитки на панель візуалізації. Обмін інформаційними панелями. Оновлення даних в опублікованих звітах. Просунута аналітика в *Power BI Desktop*. Розширені теми у *Power Query*. Створення та використання параметрів. Використання візуальних елементів, створених користувачем. Реалізація геопросторового аналізу. Реалізація безпеки даних. Створення шаблонів і пакетів вмісту. Прямі запити. Використання таблиць агрегації. Реалізація потоків даних.

**Перелік навчальних робіт студентів та оцінки їх у балах з дисципліни «Технології аналізу даних»**

| **Види робіт** | **К-сть балів** |
|---|---|
| Лабораторне заняття №1. Тема: «Імпорт з різних джерел та первинна обробка даних у системі Loginom». | 3 |
| Лабораторне заняття №2. Тема: «Оцінка якості даних». | 3 |
| Лабораторне заняття №3. Тема: «Очищення даних». | 3 |
| Лабораторне заняття №4. Тема: «Трансформація даних». | 3 |
| Лабораторне заняття №5. Тема: «Асоціативні правила у стимулюванні оптових покупців». | 3 |
| Лабораторне заняття №6. Тема: «Асоціативні правила в стимулюванні роздрібних продажів». | 3 |
| Лабораторне заняття №7. Тема: «Алгоритм k-means. Загальні принципи кластеризації». | 4 |
| Лабораторне заняття №8. Тема: «Сегментація даних на основі карт Кохонена». | 4 |
| Лабораторне заняття №9. Тема: «Регресійний аналіз даних. Лінійна і квадратична регресія» | 4 |
| Лабораторне заняття №10. Тема: «Логістична регресія. Оцінка кредитоспроможності позичальників». | 4 |
| Лабораторне заняття №11. Тема: «Імпорт даних із текстових, csv-файлів та файлів Excel». | 4 |
| Лабораторне заняття №12. Тема: «Імпорт даних із баз даних та веб-ресурсів». | 4 |
| Лабораторне заняття №13. Тема: «Створення базових візуалізацій» | 4 |
| Лабораторне заняття №14. Тема: «Публікація аналітичних звітів». | 4 |
| Модульний контроль | 20 |
| Виконання індивідуального завдання (СР) | 30 |
| **Разом: Аудиторна робота** | **70** |
| **Самостійна робота (СР)** | **30** |
| **Всього:** | **100** |

## КОНТРОЛЬ ТА КРИТЕРІЇ ОЦІНЮВАННЯ ЗНАНЬ СТУДЕНТІВ

При вивченні дисципліни використовуються наступні форми контролю знань студентів: поточний; модульний; підсумковий.

**Поточний контроль** передбачає перевірку теоретичних питань, самостійної роботи, практичних робіт та усне опитування по кожній практичній роботі. По даному виду контролю оцінювання знань здійснюється у відповідності до бального розподілу наведеного в попередній таблиці.

**Модульний контроль** передбачає виконання модульної контрольної роботи. Всі завдання оцінюються в 20 балів. Перше завдання (теоретичне) – 4 бали, друге завдання (практичне) – 8 балів, третє завдання (практичне) – 8 балів.

**Формою підсумкового контролю** є екзамен. Екзаменаційна оцінка (100 балів) є результатом виконання двох теоретичних питань (2 х 20 балів = 40 балів) та практичного завдання (60 балів).

**Результуюча оцінка з дисципліни** визначається як середня від балів набраних протягом семестру та отриманих на іспиті.

## СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

### Основний:


1. Cuesta H., Kumar S. Practical Data Analysis.Birmingham : Packt Publishing Ltd, 2016. 316 p.
2. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data /EMC Education Services. Indianapolis : John Wiley & Sons, Inc, 2015. 432 p.
3. Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of Analytical Data Models, Reports, and Dashboards. Birmingham : Packt Publishing Ltd, 2017. 802 p.
4. Roskladka A., Ivanova O., Kulazhenko V. Data Scientist: a glance into the future // Зовнішня торгівля: економіка, фінанси, право. 2019. № 3. С. 109-120