

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

СИСТЕМА УПРАВЛІННЯ ЯКІСТЮ

Система забезпечення якості освітньої діяльності та якості вищої освіти
сертифікована на відповідність ДСТУ ISO 9001:2015 / ISO 9001:2015

Кафедра цифрової економіки та системного аналізу

ЗАТВЕРДЖЕНО

вченою радою

(пост. пр. від *А. А. Мазаракі* 2020 р.)

Ректор

_____ А. А. Мазаракі



ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ / DATA ANALYSIS TECHNOLOGIES

ПРОГРАМА / COURSE SUMMARY

Київ 2020

**Розповсюдження і тиражування без офіційного дозволу КНТЕУ
заборонено**

Автор: А. А. Роскладка, доктор економічних наук, професор

Програму розглянуто і затверджено на засіданні кафедри цифрової економіки та системного аналізу 28.08.2020р., протокол № 1.

Рецензенти: В. В. Кулаженко, кандидат економічних наук,
М. Г. Шарафутдінов, бізнес-аналітик, директор з розвитку
компанії «CenterResearch&Development».

ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ/ DATAANALYSIS TECHNOLOGIES

ПРОГРАМА / COURSE SUMMARY

ВСТУП

Програма дисципліни «Технології аналізу даних» призначена для студентів бакалаврату КНТЕУ денної форми навчання галузі знань 12 «Інформаційні технології», спеціальності 124 «Системний аналіз», спеціалізацій «Системний аналіз», «Інформаційні технології та бізнес-аналітика (DataScience)».

Програму підготовлено відповідно до Стандарту вищої освіти України зі спеціальності 124 «Системний аналіз» та відповідної освітньо-професійної програми підготовки бакалаврів КНТЕУ.

Програма складається з таких розділів:

1. Мета, завдання та предмет дисципліни.
2. Передумови вивчення дисципліни як вибіркової компоненти освітньої програми.
3. Результати вивчення дисципліни.
4. Зміст дисципліни.
5. Список рекомендованих джерел.

1. МЕТА, ЗАВДАННЯ ТА ПРЕДМЕТ ДИСЦИПЛІНИ

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.

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.

Предметом вивчення дисципліни є основні положення й методи аналізу даних та їх комп'ютерна реалізація з використанням аналітичних платформ та спеціалізованих мов програмування.

2. ПЕРЕДУМОВИ ВИВЧЕННЯ ДИСЦИПЛІНИ ЯК ВИБІРКОВОЇ КОМПОНЕНТИ ОСВІТНЬОЇ ПРОГРАМИ

знання

- основ інформаційних технологій (операційна система *Windows*, бази даних, системи захисту інформації);
- основ дискретної математики, математичної логіки, алгоритмізації та програмування;
- принципів роботи із система управління базами даних;
- основ теорії ймовірностей та математичної статистики (випадкові величини та їх числові характеристики, закони розподілу випадкових величин, статистичні гіпотези та методи їх перевірки);

вміння

- вільно працювати з офісними додатками *Microsoft Word*, *Microsoft Excel*, *Microsoft PowerPoint*;
- формувати та обробляти запити до бази даних.

3. РЕЗУЛЬТАТИ ВИВЧЕННЯ ДИСЦИПЛІНИ

Дисципліна «Технології аналізу даних», як обов'язкова компонента освітньої програми, забезпечує оволодіння студентами загальними та фаховими компетентностями і досягнення ними програмних результатів навчання за освітньо-професійною програмою:

«Системний аналіз» (ОС бакалавр)

Номер в освітній програмі	Зміст компетентності	Номер теми, що розкриває зміст компетентності
<i>Загальні компетентності за освітньою програмою</i>		
ЗК3	Знання та розуміння предметної області професійної діяльності	1
ЗК5	Навички використання інформаційних і комунікаційних технологій	1-8
ЗК6	Здатність до пошуку, оброблення та аналізу інформації з різних джерел	1-8
ЗК7	Здатність до креативного та критичного мислення	5,8
ЗК8	Здатність приймати обґрунтовані рішення	2-4, 6
<i>Фахові компетентності за освітньою програмою</i>		
ФК1	Здатність аналізувати та формулювати висновки для різних типів складних управлінських задач у різних галузях народного господарства	2-4, 8
ФК2	Здатність ефективно проводити системний аналіз, здійснювати дослідження, видобування та аналіз даних з різноманітних інформаційних ресурсів на основі математичних моделей і методів науки про дані для процесів підтримки прийняття рішень	1-8
<i>Програмні результати навчання за освітньою програмою</i>		
ПРН1	Будувати концептуальну, логічну та фізичну моделі баз даних, на основі визначення особливостей зберігання даних, методів доступу тощо; розробляти таблиці реляційної бази даних, проводити нормалізацію відношень, застосовувати типи даних, обмеження і властивості елементів таблиць, виконувати операції реляційної алгебри; розробляти та адаптувати системи баз даних для різних галузей застосування	7
ПРН2	Використовувати моделі подання знань, методи добування та структурування знань, логічного виведення для розроблення баз знань та інтелектуальних систем	6, 7
ПРН3	Володіти достатніми знаннями математичних моделей і методів аналітики даних, мов моделювання та програмних засобів для виконання практичних завдань.	5-8
ПРН6	Застосовувати методи і засоби роботи з даними і знаннями, методи математичного, логіко-семантичного, об'єктного та імітаційного моделювання, технології системного і статистичного аналізу	1-6
ПРН14	Розробляти моделі і алгоритми прогнозування складних соціально-економічних процесів в умовах проектування нових інтелектуальних систем прийняття рішень за допомогою спеціалізованих пакетів програм	4, 8

«Інформаційні технології та бізнес-аналітика (DataScience)»(ОС бакалавр)

Номер в освітній програмі	Зміст компетентності	Номер теми, що розкриває зміст компетентності
<i>Загальні компетентності за освітньою програмою</i>		
K01	Здатність до абстрактного мислення, аналізу та синтезу	2-4
K07	Здатність до пошуку, оброблення та аналізу інформації з різних джерел	1-8
<i>Фахові компетентності за освітньою програмою</i>		
K22	Здатність до комп'ютерної реалізації математичних моделей реальних систем і процесів; проектувати, застосовувати і супроводжувати програмні засоби моделювання, прийняття рішень, оптимізації, обробки інформації, інтелектуального аналізу даних	2-4, 7
K23	Здатність використовувати сучасні інформаційні технології для комп'ютерної реалізації математичних моделей та прогнозування поведінки конкретних систем а саме: об'єктно-орієнтований підхід при проектуванні складних систем різної природи, прикладні математичні пакети, застосування баз даних і знань	5, 6
K28	<i>Здатність використовувати програмне забезпечення для аналізу даних (мови програмування, аналітичні платформи) з метою проведення математичних та методологічних досліджень.</i>	1-8
K29	<i>Здатність розробляти і впроваджувати моделі експертних систем і систем машинного навчання засобами комп'ютерного моделювання.</i>	2-4, 7
K30	<i>Здатність розуміти та уміло використовувати математичні та числові методи DataScience.</i>	1-4, 7, 8
<i>Програмні результати навчання за освітньою програмою</i>		
ПР11	Знати і вміти застосовувати на практиці системи управління базами даних і знань та інформаційні системи	1, 7
ПР12	Застосовувати методи і засоби роботи з даними і знаннями, методи математичного, логіко-семантичного, об'єктного та імітаційного моделювання, технології системного і статистичного аналізу	1-4
ПР13	Проектувати, реалізовувати, тестувати, впроваджувати, супроводжувати, експлуатувати програмні засоби роботи з даними і знаннями в комп'ютерних системах і мережах.	6, 8
ПР18	Володіти достатніми знаннями математичних моделей і методів аналітики даних, мов моделювання та програмних засобів для виконання практичних завдань	5-8

4. ЗМІСТ ДИСЦИПЛІНИ

Тема 1. Передобробка даних.

Історія розвитку *ArtificialIntelligence* і *BusinessIntelligence*. Характеристика фахівця з аналізу даних. *Softskillstahardskills*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 *KnowledgeDiscoveryinDatabases*. 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.

Список рекомендованих джерел

Основний: 1-4.

Додатковий: 6, 8-10, 15, 26, 27.

Інтернет-ресурси: 28-30.

Тема 2. Асоціація даних.

Афінитивний аналіз. Поняття типової транзакції. Предметний набір. Основні поняття *RulesMining*. 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.

Список рекомендованих джерел

Основний: 1-3.

Додатковий: 7, 19, 20.

Інтернет-ресурси: 28.

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.

Список рекомендованих джерел

Основний: 1- 3.

Додатковий: 5, 7, 20.

Інтернет-ресурси: 28.

Тема 4. Класифікація та регресія даних.

Огляд методів класифікації. Точність класифікації. Оцінювання класифікаційних методів.

Кореляційно-регресійний аналіз. Статистичні методи аналізу. Байєсівська 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 байєсівській 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.

Список рекомендованих джерел

Основний: 1, 3.

Додатковий: 6, 7, 11, 12, 14.

Інтернет-ресурси: 28.

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 *DataMining* у реальному часі *Real-TimeMining*. Адаптація системи до загальної концепції. Рекомендаційні машини. Класифікація рекомендаційних машин. Агентне навчання.

Проблеми аналізу даних з мережі Інтернет. Етапи *WebMining*. Категорії *WebMining*. Аналіз використання веб-ресурсів. Використання веб-структур та веб-контенту. Аналіз структури сегмента мережі. Персоналізація інформації. Пошук шаблонів в поведінці користувачів.

Список рекомендованих джерел

Основний: 1-3.

Додатковий: 5-7, 9, 11, 12, 14-18, 26.

Інтернет-ресурси: 28.

Тема 6. Інструментальні засоби аналізу даних

Програмне забезпечення в області аналізу даних. Аналітичні платформи *DeductorStudio*, *Loginom*, *RapidMiner*, *Tableau*, *Weka*, *Orange*, *NodeXL*, *Qlik*. Технології аналізу даних у продуктах *Microsoft Corporation*. Технологія моделювання даних у *Microsoft PowerPivot*. Інтерактивний інструмент *Microsoft PowerView* для дослідження, графічного відображення та представлення даних. Надбудова *Microsoft PowerQuery* в задачах бізнес-аналітики.

Хмарні технології *Microsoft* для аналізу та візуалізації даних. Організація бізнес-аналітики рівня *BusinessIntelligence (BI)*. Платформа *Microsoft BI*.

Механізм аналітики в пам'яті *xVelocity*. Налаштування *PowerBI* середовища Інтерфейс *PowerBIDesktop*. Завантаження даних *PowerBI* з різних інформаційних джерел. Імпорт даних із реляційних баз даних, текстового файлу, вхідного каналу даних та сервісів аналізу

Список рекомендованих джерел

Основний: 3, 4.

Додатковий: 5-7, 9, 11, 12, 14-19.

Інтернет-ресурси: 28-30.

Тема 7. Створення моделі даних

Зміна даних у *PowerQuery*. Трансформація, очищення та фільтрування даних у *PowerBI*. 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 *PowerBIDesktop*. 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 *PowerBI*. Зміна контексту запиту. Використання функцій фільтра у створених мірах.

Аналіз часових даних у моделі даних *PowerBI*. Створення таблиці дат. Оцінки на основі часового періоду. Зміна контексту дати. Використання функцій дати і часу. Створення напівадитивних мір.

Список рекомендованих джерел

Основний: 1-3.

Додатковий: 11, 12, 14.

Інтернет-ресурси: 28.

Тема 8. Побудова аналітичних звітів

Створення таблиць та матриць візуалізації даних у *PowerBIDesktop*. Побудова стрічкових, кругових діаграм та гістограм. Побудова лінійних та точкових діаграм. Створення візуалізацій на основі карт. Поєднання візуалізацій у *PowerBI*. Деталізація візуалізацій.

Публікація звітів та створення інформаційних панелей на порталі *PowerBI*. Опублікування файлів *PowerBIDesktopPowerBIService*. Додавання плитки на панель візуалізації. Обмін інформаційними панелями. Оновлення даних в опублікованих звітах.

Просунута аналітика в *PowerBIDesktop*. Розширені теми у *PowerQuery*. 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.

Список рекомендованих джерел

Основний: 3.

Додатковий: 11, 12, 14.

Інтернет-ресурси: 28.

5. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основний

1. Cuesta H., Kumar S. *PracticalDataAnalysis*. Birmingham : PacktPublishingLtd, 2016. 316 p.
2. *DataScience&BigDataAnalytics: Discovering, Analyzing, VisualizingandPresentingData* /EMCEducationServices. Indianapolis : JohnWiley&Sons, Inc, 2015. 432 p.
3. *Microsoft PowerBICookbook: CreatingBusinessIntelligenceSolutionsofAnalyticalDataModels, Reports, andDashboards*. Birmingham : PacktPublishingLtd, 2017. 802 p.
4. *Roskladka A., Ivanova O., Kulazhenko V. DataScientist: a glanceintothe future // Зовнішня торгівля: економіка, фінанси, право*. 2019. № 3. С. 109-120

Додатковий

5. *Гнатієнко Г. М., Снитюк В. Є. Експертні технології прийняття рішень: монографія*. Київ : Маклаут, 2008. 444 с.
6. *Матвійчук А. В. Штучний інтелект в економіці: нейронні мережі, нечітка логіка: монографія*. Київ : КНЕУ, 2011. 439 с.
7. *Олійник А. О., Субботін С. О., Олійник О. О. Інтелектуальний аналіз даних: навч. посібн.* Запоріжжя : ЗНТУ, 2012. 278 с.
8. *Adamson C. MasteringDataWarehouseAggregates: SolutionsforStarSchemaPerformance*. WileyPublishingInc., 2006. 318 p.
9. *Albright S. C., Winston W., Zappe C. DataAnalysisandDecisionMaking*. Boston : CengageLearning, 2016. 948 p.

10. Cao L., Yu P. S., Zhang C., Zhang H. Data Mining for Business Applications. Springer Science; Business Media, 2008. 402 p.
11. Coodley M. O. Introduction to Microsoft PowerBI: bring your data to life! CreateSpace Independent Publishing Platform, 2016. 128 p.
12. Etaati L. Advance Analytics with PowerBI and R. Auckland : Radacad Systems Limited, 2017. 179 p.
13. Fabrice G., Hamilton N. J. Quality Measures in Data Mining. Berlin; Heidelberg: Springer-Verlag, 2007. 361 p.
14. Ferrari A., Russo M. Introducing Microsoft PowerBI. Redmond : Microsoft Press, 2016. 407 p.
15. Han J., Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2006. 800 p.
16. Linoff G. S. Data Analysis Using SQL and Excel. Indianapolis: Wiley, 2015. 792 p.
17. Linoff G. S., Berry M. J. A. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Indianapolis: Wiley, 2011. 888 p.
18. Milton M. Head First Data Analysis: A learner's guide to big numbers, statistics, and good decisions. Sebastopol: O'Reilly Media, 2009. 435 p.
19. Rapid Miner: Data Mining Use Cases and Business Analytics Applications / Edited by Markus Hofmann & Ralf Klinkenberg. Minneapolis : CRC Press, 2004. 518 p.
20. Upton G. Categorical data analysis by example. New Jersey: John Wiley & Sons Inc, 2017. 198 p.
21. Witten I. H., Eibe F., Hall M. A. Data Mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, 2011. 630 p.
22. Лысенко Ю. Г., Н. Н. Иванов Н. Н., Минц А. Ю. Нейронные сети и генетические алгоритмы : учебн. пособ. Донецк : ООО «Юго-Восток, Лтд», 2003. 265 с.
23. Олійник А. О., Субботін С. О., Олійник О. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей : монографія. Запоріжжя : ЗНТУ, 2009. 375 с.
24. Рідкокаша А. А., Голдер К. К. Основи систем штучного інтелекту : навч. посібн. Черкаси : ВІДЛУННЯ-ПЛЮС, 2002. 240 с.
25. Руденко О. Г., Бодяньський Є. В. Штучні нейронні мережі. Харків : Компанія СМІТ, 2006. 404 с.
26. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг) : навч. посібн. Київ : КНЕУ, 2007. 376 с.
27. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібн. Запоріжжя : ЗНТУ, 2008. 341 с.

Інтернет-ресурси

28. Microsoft PowerBI Guided Learning URL: <https://docs.microsoft.com/uk-ua/power-bi/guided-learning> (дата звернення 02.10.2020).
29. Weka 3: Data Mining Software in Java. URL: <http://www.cs.waikato.ac.nz/ml/weka> (дата звернення 02.10.2020).
30. Rapid Miner Academy: learn data science and Rapid Miner from leading industry experts. URL: <https://academy.rapidminer.com/> (дата звернення 02.10.2020).

* Курсивом виділені джерела, що є у бібліотеці КНТЕУ