Міністерство освіти і науки України

Харківський національний університет імені В.Н. Каразіна

Кафедра теоретичної та прикладної системотехніки

«ЗАТВЕРДЖУЮ»

Проректор з науково-педагогічної роботи

_____ Антон ПАНТЕЛЕЙМОНОВ

"____"_____2020 р.

Робоча програма навчальної дисципліни

Аналіз даних

рівень вищої освіти перший (бакалаврський)

спеціальність 151 «Автоматизація та комп'ютерно-інтегровані технології»

освітня програма Автоматизація та комп'ютерно-інтегровані технології

вид дисципліни за вибором

факультет комп'ютерних наук

2020 / 2021 навчальний рік

Програму обговорено та рекомендовано до затвердження вченою радою факультету комп'ютерних наук
Протокол від «31» серпня 2020 р. №12

РОЗРОБНИК ПРОГРАМИ:
кандидат технічних наук, доцент, доцент кафедри теоретичної та прикладної системотехніки **Мазорчук Марія Сергіївна.**

Програму схвалено на засіданні кафедри теоретичної та прикладної системотехніки
Протокол від «31» серпня 2020 року, протокол № 1

Завідувач кафедри теоретичної та прикладної системотехніки
_____ Сергій ШМАТКОВ

Програму погоджено з гарантом освітньої програми «Автоматизація та комп'ютерно-інтегровані технології»

Гарант освітньої програми «Автоматизація та комп'ютерно-інтегровані технології»
_____ Дмитро ЛАБЕНКО

Програму погоджено методичною комісією факультету комп'ютерних наук
Протокол від «31» серпня 2020 року, протокол № 1

Голова методичної комісії факультету комп'ютерних наук
_____ Анатолій БЕРДНІКОВ

ВСТУП

Програма навчальної дисципліни «Аналіз даних» розроблена відповідно до освітньо-професійної програми підготовки першого (бакалаврського) рівня спеціальності 151 «Автоматизація та комп’ютерно-інтегровані технології».

1. Опис навчальної дисципліни

1.1. Мета викладання навчальної дисципліни.

Засвоєння студентами основ моделювання чисельних і категоріальних даних та методології статистичного аналізу, статистичного оцінювання взаємозв’язку величин, зниження розмірності даних та ін., вироблення навичок по адаптації стандартних алгоритмів до нових – чисельних рішень складних прикладних задач, а також придбання знань про пакети прикладних програм спеціального призначення і середовищ для обробки даних.

Об'єктом вивчення дисципліни «Аналіз даних» є сучасні математичні моделі та методи статистичного аналізу даних при управлінні складними комп’ютеризованими системами, у якій розробляються математичні моделі, методи й алгоритми аналізу даних, а також шляхи використання для цієї мети сучасних комп’ютерних систем і спеціалізованих пакетів прикладних програм.

Предметом вивчення є методи й алгоритми статистичного аналізу даних, оцінки їх ефективності та ін., для рішення яких розробляється спеціалізовані середовища для обробки статистичної інформації.

1.2. Основні завдання вивчення дисципліни.

- Основними завданнями вивчення навчальної дисципліни є:
- діагностування систем на основі даних моніторингу;
- статистичне оцінювання параметрів розподілів;
- аналіз категоріальних даних;
- кореляційний аналіз даних;
- множинний регресійний аналіз;
- оцінювання інформативності змінних при невизначеності даних;
- факторний аналіз і багатовимірне шкалювання.

1.3. Кількість кредитів – 3.

1.4. Загальна кількість годин – 90.

1.5. Характеристика навчальної дисципліни	
За вибором	
Денна форма навчання	Заочна (дистанційна) форма навчання
Рік підготовки	
4-й	-й
Семестр	
7-й	-й
Лекції	
24 год.	год.
Практичні, семінарські заняття	
24 год.	год.
Лабораторні заняття	
0 год.	год.
Самостійна робота	
42 год.	год.
Індивідуальні завдання	
0 год.	

1.6. Заплановані результати навчання

Відповідно до вимог освітньо-кваліфікаційного рівня підготовки за результатами вивчення дисципліни студенти повинні –

- знати методи статистичного аналізу даних в обсязі, необхідному для користування математичним апаратом та методами у галузі автоматизації, а саме:
- основні цілі та вихідні передумови застосування статистичних методів при управлінні складними комп'ютеризованими системами;
- методи попередньої обробки даних;
- основні поняття вибіркового методу;
- методи перевірки статистичних гіпотез;
- характеристики статистичного зв'язку кількісних даних;
- міри оцінювання категоріальних даних;
- методи кореляційного аналізу статистичних даних;
- методи регресійного аналізу;
- методи оцінювання інформативності змінних при невизначеності даних;
- методи факторного аналізу даних;

уміти:

- застосовувати сучасні інформаційні технології та мати навички розробляти алгоритми та комп'ютерні програми з використанням мов високого рівня для аналізу даних, створювати бази даних та використовувати інтернет-ресурси;
- використовувати різноманітне спеціалізоване програмне забезпечення для розв'язування типових інженерних задач у галузі математичного моделювання, пов'язаних з обробкою статистичної інформації;
- вибирати адекватні методи статистичного аналізу даних у відповідності з метою дослідження та характером статистичних даних;
- знаходити чисельні характеристики статистичних розподілів даних;
- перевіряти основні гіпотези щодо параметрів розподілення даних на основі вибіркових характеристик;
- аналізувати аномальні спостереження у скалярних та векторних даних;
- оцінювати зв'язок між категоріальними ознаками даних;
- знаходити рівняння лінійної регресії;
- перевіряти значущість коефіцієнтів лінійної регресії;
- знаходити довірчі інтервали для коефіцієнтів лінійної регресії;
- будувати лінійну множинну регресію;
- знаходити рівняння нелінійної регресії;
- визначати статистичні оцінки для параметрів нелінійної регресії;
- оцінювати інформативність змінних з врахуванням точності вимірювання змінних стану і наявності парної кореляції між ними;
- оцінювати довірчі інтервали для математичного очікування нелінійних залежностей;
- представляти змістовну інтерпретацію результатів статистичного аналізу;
- працювати з сучасними програмними системами і середовищами для статистичного аналізу даних (STATISTICA, SPSS, R, пакети Python);

придбати навички:

- формулювання змістовної та математичної постановок задач, здійснювання формалізації представлення даних, структуризації поставлених задач;
- засвоєння основних методів і прийомів аналізу та обробки різних видів інформації; розробки моделей та методів статистичного аналізу даних;
- проведення верифікації математичних методів, оцінки їх якості на основі перевірки існуючих статистичних гіпотез ;
- вирішення задач чисельного характеру з застосуванням спеціалізованих пакетів;

мати уявлення:

- про класичні і сучасні методи статистичного аналізу даних;

- про межу можливих застосувань методів статистичного аналізу даних.

2. Тематичний план навчальної дисципліни

Розділ 1. Описові статистики і частотні розподіли, аналіз взаємозв'язків між ознаками досліджуваних систем.

Тема 1. Аналіз даних у системах різної природи на основі моніторингу.

Застосування математичних методів у різних системах: опис, пояснення та прогнозування статистичних процесів та явищ. Основні напрямки використання математичних методів при аналізі даних. Логіка та основні етапи аналізу даних. Статистичні методи аналізу даних та задачі, що розв'язуються з їх застосуванням. Якісні та кількісні методи аналізу даних.

Вимірювання. Поняття шкали вимірювання. Номінальні шкали, порядкові шкали, інтервальні шкали, шкали відношень: припустимі перетворення, припустимі арифметичні операції, властивості, приклади.

Тема 2. Статистичне оцінювання параметрів розподілів.

Вибірка та популяція. Статистичні оцінки. Таблиця одновимірного розподілу. Відсутні значення. Частота. Частка. Варіаційний ряд. Кумулятивна частота. Щільність та відносна щільність розподілу. Крива розподілу. Класифікація розподілу за формою кривої розподілу.

Міри центральної тенденції. Середнє арифметичне. Властивості середнього арифметичного. Медіана. Мода. Квантилі: квартилі, децилі, процентилі.

Міри варіації. Варіаційний розмах. Середнє абсолютне відхилення. Дисперсія. Стандартне відхилення. Властивості дисперсії та стандартного відхилення. Порівняння варіації двох ознак. Коефіцієнт варіації. Напівквартильне відхилення. Індекс якісної варіації. Асиметрія. Ексцес. Стандартизація даних.

Тема 3. Кореляційний аналіз даних та оцінка зв'язку між різними типами даних.

Функціональний та кореляційний зв'язки. Діаграма розсіяння. Лінія регресії. Основні властивості статистичної залежності: тип, форма та тіснота зв'язку. Лінійний та нелінійний зв'язки. Прямий та зворотній зв'язки.

Коваріація. Коефіцієнт парної кореляції Пірсона: формула, область значень, умови застосування, інтерпретація. Типові помилки у використанні та інтерпретації коефіцієнту кореляції. Хибна кореляція та хибна незалежність ознак. Причинність та кореляція. Кореляційне відношення: формула, область значень, інтерпретація.

Коефіцієнти зв'язку для двох ознак, побудованих для різних типов шкал: коефіцієнт, точечний бісеріальний коефіцієнт кореляції, коефіцієнт рангової кореляції Спірмена, коефіцієнт тау Кенделла. Поняття рангу. Рангова кореляція. Коефіцієнт рангової кореляції Спірмена: коваріація рангів, виведення формули, область значень, інтерпретація, зв'язок з коефіцієнтом кореляції Пірсона. Коефіцієнти рангової кореляції Кендела: правила обчислення, область значень та інтерпретація. Порівняння коефіцієнтів Спірмена та Кендела.

Тема 4. Аналіз категоріальних даних.

Двовимірна таблиця як інструмент вивчення взаємозв'язку двох ознак: структура та правила читання. Маргінальний стовпчик та маргінальний рядок. Поняття статистичного зв'язку. Зв'язок між двома ознаками як відмінність від статистичної незалежності. Обчислення коефіцієнту хі-2 Пірсона та його інтерпретація. Число ступенів свободи. Застосування хі-2 для перевірки гіпотези про наявність зв'язку між двома ознаками. Рівень значущості. Стандартні рівні значущості в соціальних науках. Застосування хі-2 для оцінювання відповідності емпіричного і теоретичного розподілів та для порівняння двох емпіричних розподілів. Необхідні для застосування критерія хі-2 умови.

Коефіцієнти зв'язку для двох ознак, побудовані на основі хі-2 : коефіцієнт середньої квадратичної спряженості Пірсона, коефіцієнт Чупрова, коефіцієнт Крамера. Виведення формул, правила обчислення, області значень та інтерпретація. Порівняння різних

коефіцієнтів. Коефіцієнти асоціації та контингенції для таблиць розміру 2*2. Недоліки коефіцієнтів, що побудовані на основі хі-2.

Розділ 2. Формальні математичні моделі систем.

Тема 5. Множинний регресійний аналіз.

Регресійний аналіз. Основні методи побудови регресійних моделей. Модель лінійної регресії. Оцінка повноти, адекватності моделі, інтерпретація та оцінки коефіцієнтів рівнянь регресії, рівень значущості коефіцієнтів. Обмеження регресійної моделі - мультіколінеарность, гомоскедастічность.

Модель логістичної регресії. Інтерпретація коефіцієнтів логістичної регресії.

Некоректно поставлені завдання. Алгоритми, що регулярізують (робастні алгоритми): адаптивні, інваріантні. Методи регулярізації в задачах ідентифікації, апроксимації даних та прогнозування часових рядів.

Тема 6. Перевірка статистичних гіпотез. Методи оцінювання інформативності (значущості) змінних при невизначеності даних.

Поняття статистичної гіпотези. Нульова та альтернативна гіпотези. Помилка першого роду, рівень значущості та критична область. Загальна процедура перевірки статистичних гіпотез. Помилка другого роду. Однобічний та двобічний критерії. Залежні та незалежні вибірки.

Перевірка гіпотез про рівність часток (відсотків), середніх, дисперсій та коефіцієнтів кореляції у двох вибірках. Перевірка гіпотез про значущість коефіцієнтів кореляції (Пірсона, Спірмена, Кендела та коефіцієнтів, побудованих на основі хі-2).

Методи оцінюванні диференціальної інформативності з врахування точності вимірювання змінніх стану и наявність парної кореляції между ними: кореляційного аналізу, дисперсійного аналізу і методи розпізнавання образів. Статистичні оцінки довірчих інтервалів для математичного очікування нелінійних залежностей методом Монте-Карло.

Тема 7. Факторний аналіз.

Головна ідея методів факторного аналізу даних. Задачі, що розв'язуються із застосуванням факторного аналізу. Основні поняття факторного аналізу: фактор, факторне навантаження, факторне значення, загальність та специфічність.

Попередній аналіз кореляційної матриці. Вимоги до даних для факторного аналізу. Лінійна модель факторного аналізу. Інтерпретація факторних навантажень. Оцінювання загальності. Методи факторного аналізу: головних компонентів, головних факторів, максимальної правдоподібності. Принцип простої структури Терстоуна. Методи обертання факторних рішень: варімакс, квартімакс, облімін.

Порівняльний аналіз різних методів факторного аналізу. Розвідувальний та конфірматорний факторний аналіз. Факторний аналіз якісних (категоризованих) даних.

3. Структура навчальної дисципліни

Назви розділів і тем	Кількість годин											
	денна форма						заочна форма					
	усього	у тому числі					усього	у тому числі				
		л	п	лаб.	інд.	с. р.		л	п	лаб.	інд.	с. р.
1	2	3	4	5	6	7	8	9	10	11	12	13
Розділ 1. Описові статистики і частотні розподіли, аналіз взаємозв’язків між ознаками досліджуваних систем..												
Тема 1. Описові статистики і частотні розподіли, аналіз	12	2	2			8						

взаємозв'язків між ознаками досліджуваних систем												
Тема 2. Статистичне оцінювання параметрів розподілів	14	4	4			6						
Тема 3. Кореляційний аналіз даних та оцінка зв'язку між різними типами даних	14	4	4			6						
Тема 4. Аналіз категоріальних даних	14	4	4			6						
Разом за розділом 1	54	14	14			26						
Розділ 2. Формальні математичні моделі систем.												
Тема 5. Множинний регресійний аналіз	14	4	4			6						
Тема 6. Перевірка статистичних гіпотез. Методи оцінювання інформативності (значущості) змінних при невизначеності даних	8	2	2			4						
Тема 7. Факторний аналіз	14	4	2			6						
Разом за розділом 2	36	10	8			16						
Контрольна робота за розділами 1,2			2									
Усього годин	90	24	24			42						

4. Теми практичних занять

№ п/п	Назва теми	Кількість годин
1	Первинна обробка даних	2
2	Перевірка гіпотези про вид розподілу ознаки	2
3	Кореляційний аналіз даних	4
4	Аналіз категоріальних даних	2
5	Статистичне оцінювання параметрів розподілів	2
6	Множинний регресійний аналіз(лінійна форма регресійної залежності)	2
7	Множинний регресійний аналіз(поліноміальна форма регресійної залежності)	2

8	Перевірка статистичних гіпотез. Оцінювання інформативності (значущості) змінних при невизначеності даних.	4
9	Факторний аналіз	2
	Разом	22

5. Завдання для самостійної роботи

№ п/п	Зміст	Кількість годин
1	Дослідити перевірку значущості коефіцієнтів рівняння множинної регресії	6
2	Дослідити побудову довірчого та предикативного інтервалів для значення функції регресії	6
3	Провести розрахунки для двофакторного аналізу категоріальних даних	5
4	Дослідити методи оцінки рангової кореляції	5
5	Дослідити метод головних компонент	5
6	Дослідити методи факторного аналізу даних, які базуються не на методі головних компонент	5
7	Дослідити методи реалізації основних функцій аналізу даних у різних середовищах обробки даних	5
8	Підготовка до підсумкової контрольної роботи	5
	Разом	42

6. Індивідуальні завдання

Індивідуальне завдання пов'язане із застосуванням математичних моделей та методів статистичного аналізу даних в конкретному завданні, розробкою програми для його реалізації і обґрунтуванням корисності і ефективності прийнятого рішення.

Індивідуальне завдання виконується у вигляді контрольної роботи.

7. Методи навчання

Як правило, лекційні та практичні заняття проводяться аудиторно. В умовах дії карантину заняття проводяться відповідно до Наказу ректора Харківського національного університету імені В.Н. Каразіна (аудиторно або дистанційно за допомогою платформ Google Meet або Zoom).

8. Методи контролю

Контроль роботи студентів при вивченні дисципліни і засвоєння ними навчального матеріалу здійснюється на практичному зайнятті шляхом проведення «летючок», контрольних опитувань і захисту звітів по індивідуальних завданнях. Підсумковий контроль здійснюється при виконанні контрольної роботи і на заліку.

Студенти, що не захистили впродовж семестру звіти з практичних завдань, до заліку не допускаються.

Заліковий квиток містить два теоретичних і одне практичне питання. Максимальна кількість балів за відповіді на кожне теоретичне питання складає по 12 балів, на практичне питання - 16 балів. Проведення поточного контролю, письмового модульного контролю, фінальний контроль у вигляді заліку.

9. Схема нарахування балів

Поточний контроль, самостійна робота, індивідуальні завдання										Залікова робота	Сума
Розділ 1				Розділ 2			Контрольна робота, передбачена навчальним планом	Індивідуальне завдання	Разом		
Т1	Т2	Т3	Т4	Т5	Т6	Т7					
5	10	10	5	10	5	5	10		60	40	100

Т1, Т2 ... – теми розділів.

За темою Т1 студент отримує 5 балів за виконання практичної роботи 1.

За темою Т2 студент отримує по 5 балів за виконання практичних робіт 2, 3.

За темою Т3 студент отримує по 5 балів за виконання практичних робіт 4, 5.

За темою Т4 студент отримує 5 балів за виконання практичної роботи 6.

За темою Т5 студент отримує по 5 балів за виконання практичних робіт 7, 8.

За темою Т6 студент отримує 5 балів за виконання практичної роботи 9.

За темою Т7 студент отримує 5 балів за виконання практичної роботи 10.

Критерії оцінювання знань студентів за практичні роботи

Вимоги	Кількість балів
.▪ Завдання відзначається повнотою виконання без допомоги викладача. ▪ Визначає рівень поінформованості, потрібний для прийняття рішень. Вибирає інформаційні джерела,. ▪ Робить висновки і приймає рішення у ситуації невизначеності. Володіє уміннями творчо-пошукової діяльності.	5
▪ Завдання – повні, з деякими огріхами, виконані без допомоги викладача. ▪ Планує інформаційний пошук; володіє способами систематизації інформації; ▪ Робить висновки і приймає рішення у ситуації невизначеності. Володіє уміннями творчо-пошукової діяльності.	4
▪ Завдання відзначається неповнотою виконання без допомоги викладача. ▪ Студент може зіставити, узагальнити, систематизувати інформацію під керівництвом викладача; вільно застосовує вивчений матеріал у стандартних ситуаціях.	3
▪ Завдання відзначається неповнотою виконання за консультацією викладача.	2

▪ Застосовує запропонований вчителем спосіб отримання інформації, має фрагментарні навички в роботі з підручником, науковими джерелами; ▪ Вибирає відомі способи дій для виконання фахових методичних завдань.	
Завдання відзначається фрагментарністю виконання за консультацією викладача або під його керівництвом.	1

Критерії оцінювання знань студентів за контрольну роботу

Вимоги	Кількість балів
Повнота виконання завдання повна, студент здатен формулювати закони та закономірності, структурувати судження, умовиводи, доводи, описи.	8-10
Повнота виконання завдання повна, студент здатен формулювати операції, правила, алгоритми, правила визначення понять.	5-7
Повнота виконання завдання елементарна, студент здатен вибирати відомі способи дій для виконання фахових завдань.	3-5
Повнота виконання завдання фрагментарна.	1-2

Критерії оцінювання залікових робіт студентів

Вимоги	Кількість балів
Показані всебічні систематичні знання та розуміння навчального матеріалу; безпомилково виконані завдання.	35-40
Показані повні знання навчального матеріалу; помилки, якщо вони є, не носять принципового характеру.	30-35
Показано повне знання необхідного навчального матеріалу, але допущені помилки.	20-30
Показано повне знання необхідного навчального матеріалу, але допущені суттєві помилки	10-20
Показано недосконале знання навчального матеріалу, допущені суттєві помилки.	5-10
Показано недосконале знання навчального матеріалу, допущені суттєві помилки, які носять принциповий характер; обсяг знань не дозволяє засвоїти предмет.	1-5

Шкала оцінювання

Сума балів за всі види навчальної діяльності протягом семестру	Оцінка	
	для чотирирівневої шкали оцінювання	для дворівневої шкали оцінювання
90 – 100	відмінно	зараховано
70-89	добре	
50-69	задовільно	
1-49	незадовільно	не зараховано

10. Рекомендована література

Основна література

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии. – М.: ЮНИТИ, 1998. – 1000 с.
2. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с.
3. Field, Andy (2009). Discovering Statistics Using SPSS. – 855 p.
4. Field, Andy (2012). Discovering Statistics Using R. – 995 p.
5. Р. И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.

Допоміжна література

1. Анализ данных/ Н.С. Бакуменко, О.С. Радивоненко. – Учеб. пособие по лабораторному практикуму.–Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т.», 2007.– 87 с.
2. Решение экономических задач с использованием статистических пакетов обработки данных: учеб. Пособие по лаб. практикуму / М.С. Мазорчук, Н.С. Бакуменко. – Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т», 2008. – 95 с.
3. Анализ данных результатов статистических исследований в среде R [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк, Н.С. Бакуменко. – Х.: Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2018. – 36 с.
4. Статистический анализ данных в социологии [Текст] : учеб. пособие / М. С. Мазорчук, В. С. Добряк. – Х. : Нац. аэрокосм. ун-т им. Н.Е. Жуковского «Харьк. авиац. ин-т», 2012. – 36 с.

11. Посилання на інформаційні ресурси в Інтернеті, відео-лекції, інше методичне забезпечення

1. https://stepik.org/course/129/syllabus
2. https://stepik.org/course/67/syllabus
3. https://ru.coursera.org/learn/data-analytics-business
4. https://www.udemy.com/topic/python/