МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій

Затверджую
В.о.декана факультету комп'ютерних інформаційних технологій

_____ Ігор Якименко
"____"_____ 2023р.

Затверджую
В.о. проректора з науково-педагогічної роботи

_____ Віктор ОСТРОВЕРХОВ
"____"_____ 2023 р..

Затверджую
Директор ННІПІОТ

_____ Святослав Питель
"____"_____2023р.

РОБОЧА ПРОГРАМА

з дисципліни

«Прикладний аналіз даних»

Ступінь вищої освіти – перший(бакалаврський)
Галузь знань: 12 Інформаційні технології
Спеціальність: 124 Системний аналіз
Освітньо-професійна програма «Системний аналіз»

Кафедра економічної кібернетики та інформатики

Форма навчання	Курс	Семестр	Лекції	Практ	Лаб	ІРС	КПІЗ	СРС	Разом	Екзамен,
Денна	III	4	28	42		4	10	96	180	4
Заочна	III	5	8	4		-	-	168	180	5

Тернопіль 2023

Робоча програма складена на основі освітньо-професійної програми підготовки бакалавра галузі знань 12 Інформаційні технології спеціальності 124 Системний аналіз, затвердженої на засіданні вченої ради ЗУНУ (протокол №9 від 26.05.2021 р.).

Робочу програму склав: професор кафедри економічної кібернетики та інформатики Пасічник Роман Мирославович

Робоча програма затверджена на засіданні кафедри економічної кібернетики та інформатики, протокол № 1 від 28.08.2023 р.

Завідувач кафедри проф. БУЯК Леся Михайлівна

Розглянуто та схвалено групою забезпечення спеціальності системний аналіз, протокол №1 від 30.08.2023 р.

Голова ГЗС проф. ПАСІЧНИК Р.М.

Гарант ОПП проф. ПАСІЧНИК Р.М.

СТРУКТУРА РОБОЧОЇ ПРОГРАМИ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

«Прикладний аналіз даних»

1. Опис дисципліни «**Прикладний аналіз даних**»

Дисципліна – Прикладний аналіз даних	Галузь знань, напрям підготовки, освітньо-кваліфікаційний рівень	Характеристика навчальної дисципліни
Кількість кредитів ECTS 6	**Галузь знань –** 12 «Інформаційні технології»	Нормативна дисципліна, мова навчання - *українська*
Кількість залікових модулів - 3	**Спеціальність –** 124 «Системний аналіз»,	*Денна:* Рік підготовки:3 Семестр – 5 *Заочна:* Рік підготовки:3 Семестр – 6
Кількість змістових модулів - 2	Ступінь вищої освіти – бакалавр	*Денна:* лекції – 28 год.; практ – 42 год *Заочна:* лекції – 8 год.; лабор.- 4 год
Загальна кількість годин - - 180		Самостійна робота: 96 *год.,* ((КПІЗ) – 10 год.) Індивідуальна робота : *4 год.*
Тижневих годин: 12 год., з них аудиторних – 5 год		Вид підсумкового контролю – *екзамен*

2. Мета й завдання вивчення дисципліни "Прикладний аналіз даних"

2.1. Мета вивчення дисципліни

Метою викладання дисципліни "Прикладний аналіз даних" є ознайомлення студентів з методологією підтримки прийняття рішень на основі статистичних методів із застосуванням їх для розв'язання прикладних задач. .

2.2. Завдання вивчення дисципліни

В результаті вивчення курсу "Прикладний аналіз даних" студенти повинні:

- знати основні поняття попереднього аналізу вибірок, базові структури даних середовища Python, пакетів Pandas та R, метод ієрархічного кластерного аналізу, метод к-середніх, основні статистичні гіпотези, методи відсіву вибірок, методи перевірки однорідності вибірок, основи лінійної та нелінійної регресій, основи прогнозування часових рядів;
- вміти здійснювати попередній аналіз вибірок, реалізовувати кластеризацію вибірок, перевіряти гіпотези однорідності, корельованості, нормальності вибірок, будувати регресійні залежності прогнозувати часові ряди.

2.3. Найменування та опис компетентностей, формування котрих забезпечує вивчення дисциплін:

1. Здатність проводити обчислювальні експерименти, порівнювати результати експериментальних даних і отриманих рішень.
2. Здатність до аналізу, синтезу і оптимізації інформаційних систем та технологій з використанням математичних моделей і методів.
3. Уміння здійснювати комплексний статистичний аналіз і прогнозування фізичних та соціально-економічних процесів

2.4. Передумови для вивчення дисципліни.

Математичний аналіз, теорія імовірностей та математична статистика, програмування на Python.

2.5. Результати навчання:

1. Демонструвати знання сучасного рівня технологій інформаційних систем, практичні навички програмування та використання прикладних і спеціалізованих комп'ютерних систем та середовищ з метою їх запровадження у професійній діяльності.
2. Обґрунтовувати вибір технічної структури та розробляти відповідне програмне забезпечення, що входить до складу інформаційних систем та технологій.

2.6. Завдання лекційних занять

Мета проведення лекцій полягає у тому, щоб ознайомити студентів із головними питаннями курсу "Прикладний аналіз даних".

Завдання проведення лекцій полягає у:

- викладенні студентам у відповідності з програмою та робочим планом основних питань курсу "Прикладний аналіз даних";
- сформуванні у студентів цілісної системи теоретичних знань з курсу "Прикладний аналіз даних".

2.7. Завдання проведення практичних занять

Мета проведення практичних занять полягає у тому, щоб виробити у студентів практичні навички використання теоретичного матеріалу.

Завдання проведення практичних занять полягає у глибшому засвоєнні та

закріпленні теоретичних знань, одержаних на лекціях.

3. Програма дисципліни
"Прикладний аналіз даних"

Змістовий модуль 1 – Попередній аналіз та класифікація підвибірок

Тема 1. Прикладні пакети аналізу даних.

Основні завдання та інструменти аналізу даних. R та Python переваги, недоліки, рекомендовані сфери застосування. Середовище IDLE розроблення програм на Python. Пакет статистичного аналізу Pandas. Базові структури даних в Python списки, кортежі словники. Формування структур та вибір елементів із них. Методи сортування словників по ключах та значеннях. Пакет R та середовище аналітика RStudio

Тема 2. Структури даних та їх обробка в пакеті Pandas

Методи роботи із типом даних Series, його індексами та значеннями. Групова вибірка та присвоювання. Переіндексування та фільтрування. Тип даних DataFrame. Доступ назвою та за індексом стовпчика. Фільтрування за бульовими масивами. Читання та запис даних у csv файли. Групування та агрегування даних. Приклад аналізу даних звіту про долю пасажирів лайнера “Титанік”

Тема 3. Загальний аналіз вибірок із пакетом Pandas

Поняття випадковості та випадкової події. Імовірність настання події. Дискретні та неперервні випадкові величини. Закон та функція розподілу випадкової величини. Щільність розподілу випадкової величини. Рівномірно та нормально розподілені випадкові величини. Математичне сподівання, дисперсія, стандартне відхилення. Логарифмічно нормальні розподіли.

.

Тема 4. Попередній аналіз даних із пакетом Pandas

Генеральна сукупність та вибірка. Варіаційний ряд. Гістограма. Ядрова оцінка щільності розподілу. Коробкова діаграма. Її елементи та принцип побудови. Порівняння розподілів та обсягів кількох вибірок. Проведення попереднього аналізу за допомогою пакету Pandas.

Тема 5. Загальний аналіз вибірок із пакетом R

Встановлення робочого каталогу. Зчитування даних в оперативну пам’ять. Основні типи даних в R. Побудова звичайної та нормованої гістограми в R. Побудова гістограми нормального розподілу. Інсталяція та підключення пакету sm. Побудова ядрової оцінки щільності розподілу.

Побудова коробкових діаграм. Інсталяція та підключення пакету plyr. Оцінка обсягів груп за категорями. Побудова характеристик типового споживача. Середнє значення, медіана та усічене середнє. Роль характеристик розсіювання у побудові довірчих інтервалів для прогнозів. Дисперсія, середньоквадратичне розсіювання та міжквартильний розмах.

Тема 6. Задача класифікації підвибірок

Додавання ознак для класифікації. Стовпчикові діаграми. Задача ієрархічного кластерного аналізу. Метод к-середніх. Методи агрегації даних шляхом об’єднання спостережень в групи. Оцифрування ознак. Геометрична

інтерпретація подібності за відстанню. Типові відстані: евклідова, квадрат-евклідова, блок-манхетен. Типові випадки використання віддалей. Критичність розбіжностей окремих координат як критерій вибору метрик евкліда та манхеттена.

Тема 7. Ієрархічний кластерний аналіз

Віддаль між кластерами. Центроїдний метод та його недоліки. Методи далекого та найближчого сусіда. Віддаль Соренсона. Рекомендовані методи попереднього кластерного аналізу. Метод покрокового агрегування кластерів. Дендрограми та алгоритми їх побудови. Критерій зупинки. Насипи (лікті) для масштабних кластеризацій.

Відбір змінних кластеризації. Сурогатні змінні. Роль стандартизації змінних. Методи стандартизації. Стандартизація за розмахом та за допомогою z-міток. Інтерпретація результатів кластерного аналізу. Обмеженість на розмірності в методі ієрархічного кластерного аналізу. Приклади використання кластерного аналізу. Процедура NBClust.

Тема 8. Метод к-середніх.

Етапи побудови моделі. Експертне задання центрів кластерів. Ініціалізація датчиків випадкових чисел. Обмеження кількості змін центрів кластерів. Кількість проведення процедур кластеризації. Вибір найкращої кластеризації за критерієм якості. Допустимість мінімальної зміни центрів кластерів. Метод голосування кількості кластерів. Середнє значення показників по кластеру. Інтерпретація типових представників кластеру.

Змістовий модуль 2 – Перевірка статистичних гіпотез та регресійний аналіз

Тема 9. Основні статистичні гіпотези.

Гіпотеза погодження. Гіпотеза нормальності для економії кількості вимірювань. Гіпотеза експоненціальності для часу безвідмовної роботи. Гіпотези однорідності про рівність розподілів після виконання певних дій на покращення. Гіпотези незалежності. Гіпотеза про параметр розподілу. (л5)

Помилки першого та другого роду. Рівень значимості. Критичні значення. Статистичний критерій. Перевірка статистичних гіпотез про вигляд розподілу. Критерій Шапіро

Тема 10. Прогноз на основі лінійної регресійної моделі.

Аналіз наявності тренду та сезонності. Типи сезонностей. Мінливість характеру ряду. Відсікання минулих історій. Аналіз викидів. Ідентифікація та заміна викидів. Логарифмування даних із мультиплікативною сезонністю. Нормалізація часу спостережень. Індикатори періодів сезонності. Стовпчик підтримки вільних членів. Навчання моделі. Побудова прогнозу. Перехід до оригінальних значень.

Проблема перенавчання. Підбір структури моделі. Виділення навчальної та контрольної вибірки. Метод валідації. Сфери застосування лінійної регресії. Випадок множинної сезонності. Аналіз коротких сезонних часових рядів.

4. Структура залікового кредиту дисципліни "Прикладний аналіз даних"

денна

		Кількість годин				
	Лекції	Лабораторні заняття	Самостійна робота	Індивід робота	Тренінг та КПІЗ	Контрол заходи
Змістовий модуль 1 – Попередній аналіз та класифікація підвибірок						
Тема 1. Прикладні пакети аналізу даних	2	4	9		5	поточне опит.
Тема 2. Структури даних та їх обробка в пакеті Pandas	2	4	9	1		поточне опит.
Тема 3. Загальний аналіз вибірок із пакетом Pandas	3	4	9			поточне опит.
Тема 4. Попередній аналіз даних із пакетом Pandas	3	4	9	1		поточне опит.
Тема 5. Загальний аналіз вибірок із пакетом R	3	4	10			модульн контр
Змістовий модуль 2 – Перевірка статистичних гіпотез та регресійний аналіз						
Тема 6. Задача класифікації підвибірок	3	4	10		5	поточне опит.
Тема 7. Ієрархічний кластерний аналіз	3	4	10	1		поточне опит.
Тема 8. Метод к-середніх	3	4	10			поточне опит.
Тема 9. Основні статистичні гіпотези	3	5	10			поточне опит.
Тема 10. Прогноз на основі лінійної регресійної моделі	3	5	10	1		ректорс. контр.
Разом	28	42	96	4	10	180

заочна

	Кількість годин		
	Лекції	Лабораторні заняття	Самостійна робота
Тема 1. Прикладні пакети аналізу даних	1		16
Тема 2. Структури даних та їх обробка в пакеті Pandas	1		16
Тема 3. Загальний аналіз вибірок із пакетом Pandas	-	1	17
Тема 4. Попередній аналіз даних із пакетом Pandas	-	1	17
Тема 5. Загальний аналіз вибірок із пакетом R	1		17
Змістовий модуль 2 – Перевірка статистичних гіпотез та регресійний аналіз			
Тема 6. Задача класифікації підвибірок	1		10
Тема 7. Ієрархічний кластерний аналіз	1	1	10
Тема 8. Метод к-середніх	1	1	10
Тема 9. Основні статистичні гіпотези	1		10
Тема 10. Прогноз на основі лінійної регресійної моделі	1		10
Разом	8	4	168

5. Тематика практичних занять

Практичне заняття 1. Прикладні пакети аналізу даних

1.Основні завдання та інструменти аналізу даних.
2. R та Python переваги, недоліки, рекомендовані сфери застосування.
3. Середовище IDLE розроблення програм на Python.
4. Пакет статистичного аналізу Pandas.

Практичне заняття 2. Структури даних та їх обробка в пакеті Pandas

1.Методи роботи із типом даних Series, його індексами та значеннями.
2. Групова вибірка та присвоювання.
3. Тип даних DataFrame.
4. Читання та запис даних у csv файли.

Практичне заняття 3. Загальний аналіз вибірок із пакетом Pandas

1. Дискретні та неперервні випадкові величини.
2. Закон та функція розподілу випадкової величини. Щільність розподілу.
3. Рівномірно та нормально розподілені випадкові величини.
4. Логарифмічно нормальні розподіли.

Практичне заняття 4. Загальний аналіз вибірок із пакетом Pandas

1.Генеральна сукупність та вибірка. Варіаційний ряд.

2. Гістограма. Ядрова оцінка щільності розподілу.
3. Коробкова діаграма. Її елементи та принцип побудови.
4. Порівняння розподілів та обсягів кількох вибірок.

Практичне заняття 5. Загальний аналіз вибірок із пакетом R

1. Зчитування даних в оперативну пам'ять.
2. Побудова звичайної та нормованої гістограми в R.
3. Побудова коробкових діаграм.
4. Побудова характеристик типового споживача.

Практичне заняття 6. Задача класифікації підвибірок

1. Додавання ознак для класифікації.
2. Задача ієрархічного кластерного аналізу.
3. Метод к-середніх.
4. Типові відстані: евклідова, квадрат-евклідова, блок-манхетен.

Практичне заняття 7. Ієрархічний кластерний аналіз

1. Центроїдний метод . його недоліки.
2. Методи далекого та найближчого сусіда.
3. Дендрограми та алгоритми їх побудови. Критерій зупинки. Насипи (лікті) для масштабних кластеризацій.
4. Інтерпретація результатів кластерного аналізу.

Практичне заняття 8. Метод к-середніх

1. Експертне задання центрів кластерів.
2. Кількість проведення процедур кластеризації.
3. Вибір найкращої кластеризації за критерієм якості.
4. Метод голосування кількості кластерів.

Практичне заняття 9. Основні статистичні гіпотези

1. Гіпотеза погодження.
2. Гіпотеза нормальності для економії кількості вимірювань.
3. Гіпотеза експоненціальності для часу безвідмовної роботи.
4. Гіпотези однорідності про рівність розподілів після виконання певних дій на покращення.

Практичне заняття 10. Прогноз на основі лінійної регресійної моделі

1. Аналіз наявності тренду та сезонності.
2. Аналіз викидів. Ідентифікація та заміна викидів.
3. Логарифмування даних із мультиплікативною сезонністю.
4. Проблема перенавчання.

6. Комплексне практичне індивідуальне завдання.

1. Попередній аналіз вибірок за допомогою статистичних пакетів.
2. Ієрархічний кластерний аналіз
3. Метод к-середніх.
4. Основні статистичні гіпотези
5. Порівняння характеристик вибірок
6. Лінійна регресія
7. Аналіз часових рядів
8. Прогнозування часових рядів

7. Самостійна робота

№ п/п	Тематика
1.	Основні завдання та інструменти аналізу даних. R та Python переваги, недоліки, рекомендовані сфери застосування.
2.	Середовище IDLE розроблення програм на Python.
3.	Пакет статистичного аналізу Pandas.
4.	Базові структури даних в Python списки, кортежі словники.
5	Методи сортування словників по ключах та значеннях. Пакет R та середовище аналітика RStudio.
6	Методи роботи із типом даних Series, його індексами та значеннями.
7	Групова вибірка та присвоювання. Переіндексування та фільтрування. Тип даних DataFrame.
8	Закон та функція розподілу випадкової величини.
9	Щільність розподілу випадкової величини.
10	Рівномірно та нормально розподілені випадкові величини
11	Математичне сподівання, дисперсія, стандартне відхилення. Логарифмічно нормальні розподіли.
12	Генеральна сукупність та вибірка. Варіаційний ряд. Гістограма. Ядрова оцінка щільності розподілу. Проведення попереднього аналізу за допомогою пакету Pandas.
13	Задача ієрархічного кластерного аналізу. Метод к-середніх. Методи агрегації даних шляхом об’єднання спостережень в групи.
14	Віддаль між кластерами. Центроїдний метод та його недоліки. Методи далекого та найближчого сусіда. Віддаль Соренсона
15	Рекомендовані методи попереднього кластерного аналізу. Насипи (лікті) для масштабних кластеризацій.
16	Етапи побудови моделі. Експертне задання центрів кластерів. Метод голосування кількості кластерів.
17	Гіпотеза про параметри розподілу. Перевірка статистичних гіпотез про вигляд розподілу. Критерій Шапіро
18	Критерій рівності математичних сподівань Стьюдента. Випадок парних вибірок. Критерій Манна-Уітні порівняння медіан.
19	Покроковий метод зменшення моделі. Найпростіша нелінійна модель регресії. Бібліотека PlynomialFeatures. Побудова множини моделей. Аналіз коефіцієнтів детермінації. Аналіз значимостей та знаків коефіцієнтів моделі. Відбір адекватних моделей.
20	Сфери застосування лінійної регресії. Випадок множинної сезонності. Аналіз коротких сезонних часових рядів.
21	Основні особливості моделей Бокса-Дженкінса та ARIMA.
22	Підбір параметрів моделі ARIMA. Аналіз залишків моделі. Q-тест Льюнга — Бокса. Підбір параметрів моделі на основі критерію Акаїки. Створення та візуалізація прогнозів. Оцінки похибок прогнозів.

8. Тренінг з дисципліни

Тематика: Дослідження залежностей у вибірках.

Порядок проведення:

5. Попереднє дослідження вибірки. Аналіз законів розподілу за категоріями за допомогою гістограм та ядрових оцінок щільності.
6. Кластеризація вибірок ієрархічно та методом к-середніх

9. Засоби оцінювання та методи демонстрування результатів навчання

- У процесі вивчення дисципліни «Прикладний аналіз даних» використовуються наступні засоби оцінювання та методи демонстрування результатів навчання:
- - поточне опитування;
- - залікове модульне тестування та опитування;
- - аналітичні звіти, реферати, есе;
- - оцінювання результатів КПІЗ;
- - розрахункові роботи;
- - ректорська контрольна робота;
- - екзамен;

- **10. Критерії, форми поточного та підсумкового контролю**

Підсумковий бал (за 100-бальною шкалою) з дисципліни "Прикладний аналіз даних" визначається як середньозважена величина, в залежності від питомої ваги кожної складової залікового кредиту:

Заліковий модуль 1	**Заліковий модуль 2**	**Заліковий модуль 3**	**Екзамен**	**Разом**
20%	20%	20%	40%	100%
1. Усне опитування під час заняття (5 теми по 10 балів = 50 балів) 2. Письмова робота = 50 балів	1. Усне опитування підчас заняття (5 тем по 10 балів = 50 балів) 2.Письмова робота = 50 балів	1.Написання та захист КПІЗ = 60 балів. 3. Виконання завдань під час тренінгу = 40 балів	1. 3 запитання по 20 балів =60 балів 2. Задача = 40 балів	

11. **Інструменти, обладнання та програмне забезпечення, використання яких передбачає навчальна дисципліна**

№	Найменування	Номер теми
1.	Персональний комп’ютер	1-10
2.	Програмне середовище Python	1-10
3	Прикладні пакети R, Pandas	1-10

РЕКОМЕНДОВАНІ ДЖЕРЕЛА ІНФОРМАЦІЇ

.

1. Wes McKinney. Python for Data Analysis. O'Reilly Media, 2013. https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Analysis.pdf
2. Йорн Гіз. Підручник із ієрархічної кластеризації та дендрограм SciPy. https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/
3. Marco Peixeiro. Time Series Forecasting in Python. https://www.methsoft.ac.cn/scipaper_files/document_files/Manning.Time.Series.Forecasting.in.Python.pdf.
4. 8host.com. Прогнозування часових рядів за допомогою ARIMA в PYTHON 3. https://www.8host.com/blog/prognozirovanie-vremennyx-ryadov-s-pomoshhyu-arima-v-python-3/
5. Joel Grus. Data Science from Scratch. O'Reilly Media. 2019. https://covid19.uthm.edu.my/wp-content/uploads/2020/04/Data-Science-from-Scratch-First-Principles-with-Python-by-Joel-Grus-z-lib.org_.epub_.pdf
6. Dirk P. Kroese, Zdravko I. Botev, Thomas Taimre, Radislav Vaisman. Data Science and Machine Learning Mathematical and Statistical Methods. 2022. https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf.
7. Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda. Applied Text Analysis with Python. O'Reilly Media. 2018. https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/