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 ННІНОТ
_____ Святослав Питель
"_____2023р.

РОБОЧА ПРОГРАМА

з дисципліни

«Аналіз даних»

Ступінь вищої освіти – другий (магістерський)
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.	ІРС	Тре-нінг, КПІЗ	СРС	Разом	Екзамен,
Денна	І	ІІ	30	15	5	4	96	150	ІІ
Заочна	І	ІІ, ІІІ	8	4	-	-	138	150	ІІІ

31.01.2023

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 №10 від 23.06.2023 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 Роман Мирославович

Гарант ОПП доц. БАБАЛА Людмила Василівна

СТРУКТУРА РОБОЧОЇ ПРОГРАМИ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

«Аналіз даних»

1. Опис дисципліни **«Аналіз даних»**

Дисципліна – Аналіз даних	Галузь знань, напрям підготовки, освітньо-кваліфікаційний рівень	Характеристика навчальної дисципліни
Кількість кредитів ECTS 5	**Галузь знань –** 12 «Інформаційні технології»	Нормативна дисципліна циклу професійної підготовки, мова навчання - *українська*
Кількість залікових модулів - 4	**Спеціальність –** 124 «Системний аналіз»,	*Денна:* Рік підготовки:1 Семестр – 1 *Заочна:* Рік підготовки:1 Семестр – 2
Кількість змістових модулів - 2	Ступінь вищої освіти – магістр	*Денна:* лекції – 30 год.; практ.- 15 год *Заочна:* лекції – 8 год.; практ.- 4 год
Загальна кількість годин - - 150		Самостійна робота: 96 *год.,* (тренінг(КПІЗ) – 4 год.) Індивідуальна робота : 5 *год.*
Тижневих годин: 10 год., з них аудиторних – 3 год		Вид підсумкового контролю – *екзамен*

2. Мета й завдання вивчення дисципліни "Аналіз даних"

2.1. Мета вивчення дисципліни

Метою викладання дисципліни "Аналіз даних" є ознайомлення студентів з методологією підтримки прийняття рішень на основі керованих та напівкерованих методів машинного навчання із застосуванням їх для розв'язання прикладних задач. .

2.2. Завдання вивчення дисципліни

В результаті вивчення курсу "Аналіз даних" студенти повинні:

- знати основні поняття лінійного та квадратичного аналізу вибірок, машин опорних векторів, стохастичного градієнтного спуску, методу найближчих сусідів, сумішей нормальних розподілів, наївних Байєсівських класифікаторів, дерев рішень, ансамблевих методів, напівконтрольованого навчання;

- вміти здійснювати лінійний та квадратичний аналіз вибірок, реалізовувати кластеризацію вибірок за допомогою машин опорних векторів, стохастичного градієнтного спуску, найближчих сусідів, наївних Байєсівських класифікаторів, , дерев рішень, будувати прогнози на основі сумішей нормальних розподілів, ансамблевих методів, напівконтрольованого навчання та ізотонічної регресії.

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, прикладний аналіз даних, 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 "Аналіз даних".

Завдання проведення лекцій полягає у:

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

Змістовий модуль 1 – Попередній аналіз вибірок

- **Тема 1. Лінійний і квадратичний дискримінантний аналіз**

Лінійний дискримінантний аналіз (LDA) із застосуванням генеративного підходу для класифікації. Моделі класів та їх коваріацій. Випадок двійкової класифікації. Гребенева регресія.

Квадратний дискримінантний аналіз (QDA) із індивідуальною коваріаційною матрицею. Межова квадратична поверхня.

Тема 2. Машини опорних векторів

Мета алгоритму опорних векторів. Розділяюча гіперплощина. М'якгкі розділювачі. Регуляризація. Перехід просторів вищої розмірності. Розділяючі гіперплощини.

Тема 3. Стохастичний градієнтний спуск

Рух за антиградієнтом. Обмеження на крок спуску. Стохастична точка спуску. Множина точок пошуку. Використання стохастичного градієнтного спуску для у лінійних моделях машинного навчання.

.

Тема 4. Метод найближчих сусідів

Класифікація точок на основі класифікації найближчих сусідів. Підбір кількості сусідів. Підбір кількості сусідів на невеликій навчальній вибірці. Вибір метрики у відстанях між сусідами. Переваги та недоліки методу найближчих сусідів.

Тема 5. Суміші нормальних розподілів

Гаусівські процеси. Постійне ядро. Ядро білого шуму. Квадратично-експоненціальне ядро. Базисна функція (RBF). Періодичне ядро. Розподіл імовірностей для нових спостережень. Прогнозування Гаусівського процесу. Якість інтерполяції та екстраполяції.

Змістовий модуль 2 – Класифікатори

Тема 6. Наївний баєсівський класифікатор

Теорема Байєса. Докази. Апостеріорна імовірність. Класифікаційна імовірність. Наївне припущення. Класова імовірність. Умовні імовірності. Припущення щодо умовних імовірностей. Види класифікаторів: нормальний, мультиноміальний, Бернуллі.

Тема 7. Дерева рішень

Структура дерева рішень: вузли, гілки, листки, функції, правила, результати. Алгоритм дерева класифікації та регресії. Сильні та слабкі сторони підходу дерева рішень

Тема 8. Ансамблеві методи

Призначення ансамблевих методів. Бутстрапінг як метод випадкового створення вибірок. Беггінг як метод пакетування із агрегацією бустрапінгу. Переваги пакетування. Моделі випадкового лісу, як удосконалений беггінг.

Тема 9. Напівконтрольоване навчання

Основна відмінність між контрольованим та неконтрольованим навчанням. Використання невеликої кількості розмічених та великої кількості нерозмічених даних. Приклади напівконтрольованого навчання. Маркування аудіофайлів, веб-контенту, білкових структур.

Тема 10. Ізотонічна регресія.

Проблема знаходження монотонної функції, що мінімізує похибку на експериментальних даних. Покрокове використання динамічного програмування. Застосування ізотонічної регресії. Корекція прогнозованих даних. Моделювання порядкових змінних. Врахування відсутніх значень. Виявлення викидів.

4. Структура залікового кредиту дисципліни "Аналіз даних"

Денна форма

	Кількість годин				
	Лекції	Практичні заняття	Самостійна робота	Індивід робота	Контрол заходи
Змістовий модуль 1 – Попередній аналіз вибірок					
Тема 1. Лінійний і квадратичний дискримінантний аналіз	3	2	10		поточне опит.
Тема 2. Машини опорних векторів	3	1	10	1	поточне опит.
Тема 3. Стохастичний градієнтний спуск	3	2	10		поточне опит.
Тема 4. Метод найближчих сусідів	3	2	9	1	поточне опит.
Тема 5. Суміші нормальних розподілів	3	1	10		модульн контр
Змістовий модуль 2 – Класифікатори					
Тема 6. Наївний баєсівський класифікатор	3	1	10	1	поточне опит.
Тема 7. Дерева рішень	3	1	9		поточне опит.
Тема 8. Ансамблеві методи	3	2	9	1	поточне опит.
Тема 9. Напівконтрольоване навчання	3	1	10		поточне опит.
Тема 10. Ізотонічна регресія	3	2	9	1	ректорс. контр.
Тренінг			4		
Разом	30	15	100	5	

Заочна форма

	Кількість годин		
	Лекції	Практичні заняття	Самостійна робота
Тема 1. Лінійний і квадратичний дискримінантний аналіз	1	-	13
Тема 2. Машини опорних векторів	1	-	13
Тема 3. Стохастичний градієнтний спуск	-	1	14
Тема 4. Метод найближчих сусідів	-	1	14
Тема 5. Суміші нормальних розподілів	1	-	14
Тема 6. Наївний баєсівський класифікатор	1		14
Тема 7. Дерева рішень	1	1	14
Тема 8. Ансамблеві методи	1	1	14
Тема 9. Напівконтрольоване навчання	1		14
Тема 10. Ізотонічна регресія	1		14
Разом	8	4	138

5. Тематика практичних занять

Практичне заняття 1. Лінійний і квадратичний дискримінантний аналіз.

1. Лінійний дискримінантний аналіз (LDA)
2. Гребенева регресія.
3. Квадратний дискримінантний аналіз (QDA) із індивідуальною коваріаційною матрицею.
4. Межова квадратична поверхня.

.

Практичне заняття 2. Машини опорних векторів

1. Мета алгоритму опорних векторів.
2. Розділяюча гіперплощина. М’ягкі розділювачі.
3. Регуляризація.
4. Перехід просторів вищої розмірності.

Практичне заняття 3. Стохастичний градієнтний спуск

1. Організація методу спуску
2. Обмеження на крок спуску.
3. Множина точок пошуку.
4. Використання стохастичного градієнтного спуску у лінійних моделях машинного навчання.

Практичне заняття 4. Метод найближчих сусідів

1. Класифікація точок на основі класифікації найближчих сусідів.
2. Підбір кількості сусідів.
3. Вибір метрики у відстанях між сусідами.
4. Переваги та недоліки методу найближчих сусідів.

Практичне заняття 5. Суміші нормальних розподілів

1. Гаусівські процеси. Постійне ядро.
2. Ядро білого шуму.
3. Квадратично-експоненціальне ядро.
4. Прогнозування Гаусівського процесу.

Практичне заняття 6. Наївний баєсівський класифікатор

1. Теорема Байєса. Докази. Апостеріорна імовірність.
2. Класифікаційна імовірність. Наївне припущення.
3. Класова імовірність. Умовні імовірності.
4. Види класифікаторів: нормальний, мультиноміальний, Бернуллі.

Практичне заняття 7. Дерева рішень

1. Структура дерева рішень.
2. Вузли, гілки, листки, функції, правила, результати.
3. Алгоритм дерева класифікації та регресії.
4. Сильні та слабкі сторони підходу дерева рішень.

Практичне заняття 8. Ансамблеві методи

1. Призначення ансамблевих методів.
2. Бутстрапінг як метод випадкового створення вибірок.
3. Беггінг як метод пакетування із агрегацією бустрапінгу.
4. Моделі випадкового лісу, як удосконалений беггінг..

Практичне заняття 9. Напівконтрольоване навчання

1. Основна відмінність між контрольованим та неконтрольованим навчанням.
2. Використання невеликої кількості розмічених та великої кількості нерозмічених даних.
3. Приклади напівконтрольованого навчання.
4. Маркування аудіофайлів, веб-контенту, білкових структур.

Практичне заняття 10. Ізотонічна регресія

1. Проблема знаходження монотонної функції, що мінімізує похибку на експериментальних даних.
2. Покрокове використання динамічного програмування.
3. Корекція прогнозованих даних.
4. Моделювання порядкових змінних.

6. Комплексне практичне індивідуальне завдання.

1. Попередній лінійний та квадратичний аналіз вибірок
2. Байєсівська класифікація
3. Класифікація деревом рішень .
4. Застосування ансамблевих методів

7. Самостійна робота

№ п/п	Тематика
1.	Лінійний дискримінантний аналіз (LDA) із застосуванням генеративного підходу для класифікації.
2.	Моделі класів та їх коваріацій. Випадок двійкової класифікації.
3.	Гребенева регресія.
4.	Квадратний дискримінантний аналіз (QDA) із індивідуальною коваріаційною матрицею.
5	Межова квадратична поверхня.
6	Мета алгоритму опорних векторів.
7	Розділяюча гіперплощина. М'ягкі розділювачі..
8	Регуляризація. Перехід до просторів вищої розмірності.
9	Розділяючі гіперплощини.
10	Рух за антиградієнтом. Обмеження на крок спуску.
11	Стохастична точка спуску. Множина точок пошуку.
12	Використання стохастичного градієнтного спуску у лінійних моделях машинного навчання.
13	Класифікація точок на основі класифікації найближчих сусідів. Підбір кількості сусідів.
14	Підбір кількості сусідів на невеликій навчальній вибірці.
15	Вибір метрики у відстанях між сусідами.
16	Переваги та недоліки методу найближчих сусідів.
17	Гаусівські процеси. Постійне ядро.
18	Ядро білого шуму.
19	Квадратично-експоненціальне ядро.
20	Базисна функція (RBF). Періодичне ядро.
21	Розподіл імовірностей для нових спостережень. Прогнозування Гаусівського процесу.
22	Прогнозування Гаусівського процесу. Якість інтерполяції та екстраполяції.
23	Проблема знаходження монотонної функції, що мінімізує похибку на експериментальних даних.
24	Покрокове використання динамічного програмування.
25	Корекція прогнозованих даних.
	Моделювання порядкових змінних.

8. Тренінг з дисципліни

Тематика: Ансамблеві методи.

Порядок проведення:

1. Обрати індивідуальні дані
2. .Використати бутстрапінг для випадкового створення вибірок.
3. Використати беггінг як метод пакетування із агрегацією бустрапінгу.
4. Використати моделі випадкового лісу, як удосконалений беггінг.

5. Порівняти ефективність методів

9. Засоби оцінювання та методи демонстрування результатів навчання

- У процесі вивчення дисципліни «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 "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 scikit-learn, numpy, scipy	1-10

РЕКОМЕНДОВАНІ ДЖЕРЕЛА ІНФОРМАЦІЇ

.

1. Scikit-learn. User Guide. https://scikit-learn.org/stable/user_guide.html
2. О. І. Шеремет,О. В. Садовой. Метод опорних векторів (SVM). https://www.dstu.dp.ua/Portal/Data/74/72/3st13-17.pdf
3. Scikit-learn. Stochastic Gradient Descent. https://scikit-learn.org/stable/modules/sgd.html
4. Scikit-learn. SGD: Maximum margin separating hyperplane. https://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_separating_hyperplane.html#sphx-glr-auto-examples-linear-model-plot-sgd-separating-hyperplane-py
5. SGD: Weighted samples. https://scikit-learn.org/stable/auto_examples/linear_model/plot_sgd_weighted_samples.html#sphx-glr-auto-examples-linear-model-plot-sgd-weighted-samples-py
6. Nearest Neighbors regression. https://scikit-learn.org/stable/auto_examples/neighbors/plot_regression.html#sphx-glr-auto-examples-neighbors-plot-regression-py
7. Nearest Neighbors. https://scikit-learn.org/stable/modules/neighbors.html
8. Kaixin Wang. Introduction to Gaussian process regression. https://medium.com/data-science-at-microsoft/introduction-to-gaussian-process-regression-part-1-the-basics-3cb79d9f155f
9. Naive Bayes Classifiers. https://www.geeksforgeeks.org/naive-bayes-classifiers/.
10. Decision Tree. https://www.geeksforgeeks.org/decision-tree//