СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

1. Загальна інформація про навчальну дисципліну

Повна назва навчальної дисципліни	Вступ до науки про дані
Повна офіційна назва закладу вищої освіти	Сумський державний університет
Повна назва структурного підрозділу	Факультет електроніки та інформаційних технологій. Кафедра комп'ютерних наук
Розробник(и)	Москаленко В`ячеслав Васильович
Рівень вищої освіти	Другий рівень вищої освіти, НРК – 7 рівень, QF-LLL – 7 рівень, FQ-EHEA – другий цикл
Семестр вивчення навчальної дисципліни	8 тижнів протягом 3-го семестру
Обсяг навчальної дисципліни	Обсяг становить 5 кред. ЄКТС, 150 год. Для денної форми навчання 64 год. становить контактна робота з викладачем (16 год. лекцій, 48 год. лабораторних занять), 86 год. становить самостійна робота.
Мова викладання	Українська

2. Місце навчальної дисципліни в освітній програмі

Статус дисципліни	Обов`язкова навчальна дисципліна для всіх освітніх програм спеціальності 122 "Комп`ютерні науки"
Передумови для вивчення дисципліни	Сховища даних
Додаткові умови	Додаткові умови відсутні
Обмеження	Обмеження відсутні

3. Мета навчальної дисципліни

Досягнення студентами сучасного конструктивного, фундаментального мислення та комплексу спеціальних знань в галузі інтелектуального аналізу даних та машинного навчання.

4. Зміст навчальної дисципліни

Тема 1 Первинний аналіз даних з Pandas Робота з векторами в бібліотеці NumPy. Робота з даними в бібліотеці Pandas.
Тема 2 Візуальний аналіз даних Основи matplotlib, seaborn та ploty. Візуальний аналіз даних. Приклади використання бібліотек візуального аналізу.

Тема 3 Дерева рішень, метод найближчого сусіда та лінійні моделі Дерева рішень та модель найближчого сусіда. Аналіз демографічних даних. Лінійні моделі. Ідентифікація користувача за допомогою логістичної регресії.
Тема 4 Композиція моделей, оцінювання якості моделей та інформативності ознак Паралельна композиція моделей. Алгоритм випадкового лісу. Вивчення метрик якості класифікаційних моделей (Precision, Recall, ROC-AUC, Accuracy, F1-score). Послідовна композиція моделей. Оцінювання інформативності ознак за допомогою градієнтного бустінгу. Робота з незбалансованими вибірками. Вивчення метрик якості регресійних моделей (MAE, MSE, RMSE).

5. Очікувані результати навчання навчальної дисципліни

Після успішного вивчення навчальної дисципліни здобувач вищої освіти зможе:

РН1	використовувати бібліотеки для маніпуляції вибірковими даними
РН2	використовувати бібліотеки для візуального аналізу вибіркових даних
РН3	розробляти та оптимізувати моделі інтелектуального аналізу даних
РН4	вимірювати продуктивність моделей аналізу даних
РН5	конструювати інформативний ознаковий опис спостережень

6. Роль навчальної дисципліни у досягненні програмних результатів

Програмні результати навчання, досягнення яких забезпечує навчальна дисципліна. Для спеціальності 122 Комп`ютерні науки:

ПР1	Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп'ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп'ютерних наук та на межі галузей знань.
ПР8	Розробляти математичні моделі та методи аналізу даних (включно з великим).
ПР9	Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).
ПР11	Створювати нові алгоритми розв'язування задач у сфері комп'ютерних наук, оцінювати їх ефективність та обмеження на їх застосування
ПР16	Виконувати дослідження у сфері комп'ютерних наук.

7. Роль освітнього компонента у формуванні соціальних навичок

Загальні компетентності та соціальні навички, формування яких забезпечує навчальна дисципліна:

СН1	Здатність до абстрактного мислення, аналізу та синтезу.
СН2	Здатність застосовувати знання у практичних ситуаціях.
СН3	Здатність вчитися й оволодівати сучасними знаннями.

СН4	Здатність генерувати нові ідеї (креативність).

8. Види навчальних занять

Тема 1. Первинний аналіз даних з Pandas
Лк1 "Робота з векторами в бібліотеці NumPy" (денна) Створення та ініціалізація багатовимірних масивів в NumPy. Основні типи даних в NumPy. Індексація в масивах та зрізи даних в NumPy.
Лк2 "Робота з даними в бібліотеці Pandas" (денна) Зчитування набору даних в дата-фрейми Pandas та їх перегляд. Індексування даних в дата-фреймах. Застосування функцій apply, map, replace. Групування даних та таблиці спряженості в Pandas. Перетворення дата-фреймів. Приклади використання бібліотеки Pandas в задачі аналізу відтоку клієнтів телекомунікаційного оператора.
Лб1 "Аналіз даних про доходи населення. Частина 1." (денна) Використання бібліотеки pandas для оброблення даних про доходи населення.
Лб2 "Аналіз даних про доходи населення. Частина 2." (денна) Використання бібліотеки pandas для аналізу даних про доходи населення.
Лб3 "Аналіз даних про пасажирів лайнеру "Титанік". Частина 1." (денна) Використання бібліотеки pandas для оброблення даних про пасажирів лайнеру.
Лб4 "Аналіз даних про пасажирів лайнеру "Титанік". Частина 2." (денна) Використання бібліотеки pandas для аналізу даних про пасажирів лайнеру.
Тема 2. Візуальний аналіз даних
Лк3 "Основи matplotlib, seaborn та ploty" (денна) Методи бібліотеки matplotlib. Методи бібліотеки seaborn. Методи бібліотеки ploty. Візуалізація статистичних характеристик і розподілу даних і агрегацій даних.
Лк4 "Приклади використання бібліотек візуального аналізу" (денна) Використання бібліотек візуалізації для аналізу даних про відтік клієнтів телекомунікаційного оператора.
Лб5 "Візуальний аналіз даних про публікації на сайті. Частина1." (денна) Використання бібліотеки matplotlib для візуального аналізу даних про публікації на сайті.
Лб6 "Візуальний аналіз даних про публікації на сайті. Частина 2." (денна) Використання бібліотек seaborn та ploty для візуального аналізу даних про публікації на сайті.

Лб7 "Візуальний аналіз даних про пасажирів Титаніку. Частина 1." (денна) Використання бібліотек візуального аналізу даних на прикладі набору даних про пасажирів лайнеру.
Лб8 "Візуальний аналіз даних про пасажирів Титаніку. Частина 2." (денна) Використання бібліотек pandas, matplotlib та seaborn для перевірки гіпотез на даних про пасажирів лайнеру.
Тема 3. Дерева рішень, метод найближчого сусіда та лінійні моделі
Лк5 "Дерева рішень та модель найближчого сусіда" (денна) Ідеї та принципи побудови дерев рішень для задач класифікації та регресії. Візуалізація дерев рішень. Ідеї та принципи класифікаційного і регресійного аналізу на основі k-найближчих сусідів. Вибір значень гіперпараметрів на основі крос-валідації. Приклад використання дерев рішень та k-найближчих сусідів для розпізнавання рукописних цифр.
Лк6 "Лінійні моделі" (денна) Лінійна регресія. Метод найменших квадратів. Метод максимальної правдоподібності. Розкладання помилок на зміщення і розкид. Регуляризація лінійної регресії. Логістична регресія і метод максимально правдоподібності. Лінійний класифікатор. L2-регуляризація логістичної функції втрат.
Лб9 "Аналіз демографічних даних. Частина 1." (денна) Використання методу дерева рішень для аналіз демографічних даних.
Лб10 "Аналіз демографічних даних. Частина 2." (денна) Використання методів найближчого сусіда та лінійних моделей для аналізу демографічних даних.
Лб11 "Ідентифікація користувача за допомогою логістичної регресії. Частина 1." (денна) Вивчення даних задачі про ідентифікацію користувача.
Лб12 "Ідентифікація користувача за допомогою логістичної регресії. Частина 2." (денна) Використання логістичної регресії для розв'язання задачі про ідентифікацію користувача.
Тема 4. Композиція моделей, оцінювання якості моделей та інформативності ознак
Лк7 "Паралельна композиція моделей та аналіз великих даних" (денна) Ансамблі моделей. Бутстреп вибірки. Бегінг. Отримання незміщенної оцінки помилки на підвибірках поза бутстреп-вибіркою (out-of-bag error). Алгоритм випадкового лісу та його параметри. Приклад використання алгоритму випадкового лісу на даних про відтік клієнтів телекомунікаційного оператора. Варіація та декореляційний ефект в паралельній композиції моделей. Переваги і недоліки паралельної композиції моделей аналізу даних. Оцінювання якості класифікаційних моделей (Precision, Recall, ROC-AUC, Accuracy, F1-score).

Лк8 "Послідовна композиція моделей" (денна) Бустінг та градієнтний бустінг. Алгоритми Adaboost та Xgboost. Приклад використання градієнтного бустінгу для аналізу даних відтоку клієнтів телекомунікаційного оператора. Оцінювання якості регресійних моделей (MAE, MSE, RMSE). Оцінювання важливості (інформативності) ознак. Оптимізація гіперпараметрів моделі XGBoost з використанням hyperopt. Оцінювання важливості (інформативності) ознак. Оптимізація гіперпараметрів моделі XGBoost з використанням hyperopt.
Лб13 "Аналіз великих наборів даних кредитного скорінгу на основі алгоритму випадкового лісу. Частина 1." (денна) Використання моделі випадкового лісу в задачі кредитного скорінгу з дефолтними значеннями гіперпараметрів
Лб14 "Аналіз великих наборів даних кредитного скорінгу на основі алгоритму випадкового лісу. Частина 2." (денна) Оптимізація гіперпараметрів моделі випадкового лісу в задачі кредитного скорінгу.
Лб15 "Вивчення метрик якості класифікації. Частина 1." (денна) Написання коду оптимізації параметрів регуляризації моделі класифікації для максимізації ROC-кривої
Лб16 "Вивчення метрик якості класифікації. Частина 2." (денна) Написання коду оптимізації гіперпараметрів моделі для обмеження частки помилок першого або другого роду.
Лб17 "Оцінювання інформативності ознак за допомогою градієнтного бустінгу. Частина 1." (денна) Оцінювання інформативності на основі алгоритму градієнтного бустінгу з параметрами за замовчуванням.
Лб18 "Оцінювання інформативності ознак за допомогою градієнтного бустінгу. Частина 2." (денна) Оптимізація гіперпараметрів моделі градієнтного бустінгу та оцінювання інформативності ознак.
Лб19 "Робота з незбалансованими вибірками. Частина 1." (денна) Аналіз незбалансованості вибірки та регулювання ваги екземплярів даних для оброблення незбалансованості класів.
Лб20 "Робота з незбалансованими вибірками. Частина 2." (денна) Практичне використання Blagging алгоритму для врахування незбалансованості навчальних даних.

Лб21 "Налаштування гіперпараметрів регресійних моделей за валідаційними метриками для оцінювання якості вина. Частина 1." (денна) Використання алгоритму випадкового лісу та лінійних моделей для оцінювання якості вина.
Лб22 "Налаштування гіперпараметрів регресійних моделей за валідаційними метриками для оцінювання якості вина. Частина 2." (денна) Оптимізація гіперпараметрів алгоритму випадкового лісу та лінійних моделей для оцінювання якості вина.
Лб23 "Прогнозування затримок вильоту літаків з використанням гібридної моделі. Частина 1." (денна) Поєднання логістичної регресії та алгоритму Xgboost для прогнозування затримок вильоту літаків.
Лб24 "Прогнозування затримок вильоту літаків з використанням гібридної моделі. Частина 2." (денна) Поєднання логістичної моделі та алгоритму sklearn.ensemble.GradientBoostingClassifier для прогнозування затримок вильоту літаків.

9. Стратегія викладання та навчання

9.1 Методи викладання та навчання

Дисципліна передбачає навчання через:

МН1	Лекційне навчання
МН2	Проблемне навчання
МН3	Практикоорієнтоване навчання

Лекції надають студентам теоретичні матеріали з тем дисципліни, що є основою для проблемного навчання здобувачів вищої освіти (РН1-РН5). Лекції доповнюються практичними заняттями, що надають студентам можливість застосовувати теоретичні знання на практичних прикладах (РН1-РН5). Проблемному навчанню сприятиме підготовка до лекцій. В особливих умовах застосовуються методи та засоби електронного навчання на базі платформи mix.sumdu.edu.ua

Під час проведення занять студенти отримують навички комунікації, вміння працювати в команді, здатність логічно і системно мислити, аргументовано висловлювати свої думки. Виконання лабораторних робіт допоможе студентам розвивати та реалізувати навички логічного та системного мислення, тайм-менеджменту, самостійного опрацювання матеріалу.

9.2 Види навчальної діяльності

НД1	Інтерактивні лекції
НД2	Підготовка до лекцій
НД3	Виконання лабораторної роботи

10. Методи та критерії оцінювання

10.1. Критерії оцінювання

Визначення	Чотирибальна національна шкала оцінювання	Рейтингова бальна шкала оцінювання
Відмінне виконання лише з незначною кількістю помилок	5 (відмінно)	$90 \leq RD \leq 100$
Вище середнього рівня з кількома помилками	4 (добре)	$82 \leq RD < 89$
Загалом правильна робота з певною кількістю помилок	4 (добре)	$74 \leq RD < 81$
Непогано, але зі значною кількістю недоліків	3 (задовільно)	$64 \leq RD < 73$
Виконання задовольняє мінімальним критеріям	3 (задовільно)	$60 \leq RD < 63$
Можливе повторне складання	2 (незадовільно)	$35 \leq RD < 59$
Необхідний повторний курс з навчальної дисципліни	2 (незадовільно)	$0 \leq RD < 34$

10.2 Методи поточного формативного оцінювання

	Характеристика	**Дедлайн, тижні**	**Зворотний зв'язок**
МФО1 Експрес-тестування	Експрес-тестування призначене для проміжного оцінювання рівня засвоєння теоретичного матеріалу. Проводиться на початку наступної лекції з використанням платформи електронного навчання MIX. Результати тестування обговорюються протягом виконання завдання.	Протягом лекційного заняття	https://mix.sumdu.edu.ua, Google meet
МФО2 Опитування та усні коментарі викладача за його результатами	Призначені для контролю засвоєння теоретичних знань поточної і минулих лекцій. Проводиться протягом дискусій і обговорень проблематики лекційного заняття.	Протягом лекційного заняття	Google meet
МФО3 Проміжне оцінювання виконання лабораторних завдань	Призначено для перевірки теоретичних та практичних знань, отриманих протягом модуля. Тестові питання та завдання для виконання рефакторингу, реінжинірингу і верифікації програмного забезпечення.	згідно графіку навчального процесу	https://mix.sumdu.edu.ua, особистий кабінет

10.3 Методи підсумкового сумативного оцінювання

	Характеристика	Дедлайн, тижні	Зворотний зв'язок
МСО1 Проміжний модульний контроль	Проміжний модульний контроль призначений для перевірки рівня засвоєння теоретичного матеріалу. Проводиться у форматі тестування засобами системи mix.sumdu.edu.ua. Оцінка за проміжний модульний контроль не перескладається.	згідно графіку навчального процесу	https://mix.sumdu.edu.ua
МСО2 Оцінювання участі в дискусії	Участь в дискусії не є обов'зковим видом завдання, але є необхідним для отримання максимальної оцінки за курс. Дискусії та обговорення направлені на отримання навичок пошуку, аналізу інформації, формулювання висновків та висловлення власної позиції щодо оголошених проблемних питань державною (чи англійською) мовою. Для отримання максимальної оцінки студент має не лише написати власний пост, а й прокоментувати, принаймі, два пости інших студентів.	В кінці лекції	Google Meet, https://mix.sumdu.edu.ua
МСО3 Оцінювання звіту за результатами виконання лабораторних робіт	Звіт за результатами виконання лабораторних робіт повинен містити ілюстрацію основних результатів виконання завдань відповідно методичних вказівок. Для успішного зарахування необхідно виконати мінімальний рівень завдання та оформити звіт відповідно вимогам до оформлення звітів. Оцінка зі звіту може бути один раз підвищення за умови усунення зауважень до представлених результатів і надсилання звіту у вказані терміни. В разі затримки термінів виконнання оцінка не може бути підвищена.	До початку наступного лабораторного заняття	Google Meet, https://mix.sumdu.edu.ua

Контрольні заходи:

	Максимальна кількість балів	Мінімальна кількість балів	Можливість перескладання з метою підвищення оцінки
3 семестр	**100 балів**		
МСО1. Проміжний модульний контроль	**20**		

	2x10	20	10	Ні
МСО2. Оцінювання участі в дискусії		**16**		
	8x2	16	8	Ні
МСО3. Оцінювання звіту за результатами виконання лабораторних робіт		**64**		
	16x4	64	20	Ні

До диференційованого заліку необхідно виконати всі лабораторні роботи на мінімальний рівень складності завдань. При успішному (отримання персоніфікованого сертифікату із вказівкою рівня успішності) вивченні масових відкритих онлайн курсів можуть бути реалізовані наступні варіанти перезарахувань частини кредитів: 1. Для курсу https://www.coursera.org/learn/python-data-analysis можуть бути перезараховані тема 1 в обсязі 15 годин/0,5 кредит (10 балів); 2. https://www.coursera.org/projects/python-for-data-visualization-seaborn може бути перезарахована тема 2 в обсязі 15 годин/0,5 кредит (10 балів); 3. Для курсу https://www.coursera.org/learn/machine-learning можуть бути перезараховані тема 3 в обсязі 15 годин/0,5 кредит (10 балів). Також викадачем можуть бути розглянуті інші масові відкриті онлайн курси за умови попереднього аналізу структури курсу.

11. Ресурсне забезпечення навчальної дисципліни

11.1 Засоби навчання

ЗН1	Бібліотечні фонди
ЗН2	Комп'ютери, комп'ютерні системи та мережи
ЗН3	Мультимедіа, відео- і звуковідтворювальна, проєкційна апаратура (відеокамери, проєктори, екрани, смартдошки тощо)
ЗН4	Програмне забезпечення (дистрибутив python, jupyter notebook або Google colab)

11.2 Інформаційне та навчально-методичне забезпечення

Основна література	
1	Моделі і методи інтелектуального аналізу багатовимірних даних за умов апріорної невизначеності : монографія / В. В. Москаленко. – Суми : Сумський державний університет, 2020. – 184 с. –https://essuir.sumdu.edu.ua/handle/123456789/77692
2	Egger, D. Data Science Math Skills [Електронний ресурс] / D. Egger, P. Bendich. — Duke University, 2020.
3	Талах М.В., Дворжак В.В. Інтелектуальний аналіз даних. Частина 1 / М.В. Талах, В.В. Дворжак – Чернівці: Технодрук, 2022. – 367 с. – https://archer.chnu.edu.ua/xmlui/handle/123456789/6751

4	https://mix.sumdu.edu.ua/textbooks/57940/index.html (посилання на платформу дистанційного навчання Mix)
Допоміжна література	
5	Intelligence Science and Big Data Engineering [Електронний ресурс] : 7th International Conference, IScIDE 2017, Dalian, China, September 22-23, 2017, Proceedings / edited by Yi Sun, Huchuan Lu, Lihe Zhang, Jian Yang, Hua Huang. — 1st ed. 2017. — Cham
6	https://scikit-learn.org/stable/tutorial/index.html (посилання на навчальні матеріали (англ. мовою) з використання бібліотеки sclearn)
7	https://mlcourse.ai/book/index.html(посилання на курс навчальних матеріалів (англ. мовою) з основ науки про дані)
8	https://lectures.scientific-python.org (курс лекційних матеріалів (англ. мовою) з використання бібліотек аналізу даних)
9	https://www.kaggle.com/ (посилання на платформу для дослідників даних різного рівня)
10	https://archive.ics.uci.edu/ (посилання на репозиторій наборів даних з різних галузей знань)