СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

ВСТУП ДО ТЕХНОЛОГІЇ DATA SCIENCE

Галузь знань	12 Інформаційні технології
Шифр та назва спеціальності	123 Комп'ютерна інженерія
Назва освітньо-професійної програми	Комп’ютерні мережі та Інтернет
Рівень вищої освіти	Перший (бакалаврський)
Факультет	Інформаційних технологій та кібербезпеки
Кафедра	Інформаційних та комп’ютерних систем
Статус навчальної дисципліни	ОК-30 ОПП «Комп’ютерні мережі та Інтернет»
Форма навчання	Денна

Викладач

Тихонова Олена Вікторівна
elena.tykhonova@suitt.edu.ua

Старший викладач кафедри інформаційних та комп’ютерних систем,
кандидат технічних наук

Загальна інформація про дисципліну

Анотація до дисципліни	Курс «Вступ до технології Data Science» спрямований на формування базових теоретичних і практичних компетентностей у галузі науки про дані. У межах курсу розглядаються основні етапи життєвого циклу Data Science-проєкту – від збирання та попередньої обробки даних до побудови, оцінювання й розгортання моделей. Особлива увага приділяється дослідницькому аналізу даних, інженерії ознак, вибору алгоритмів машинного навчання та оцінці якості моделей. Здобувачі ознайомлюються з сучасними інструментами аналізу даних і бібліотеками машинного навчання. Розглядаються задачі класифікації, регресії, аналізу текстів і часових рядів. Окремо висвітлюються питання

	MLOps та життєвого циклу моделей. Курс також охоплює етичні аспекти використання даних і штучного інтелекту.
Мета дисципліни	Сформувати у здобувачів системне уявлення про методологію та інструменти науки про дані, а також практичні навички збирання, попередньої обробки, дослідницького аналізу даних, побудови, оцінювання та впровадження моделей машинного навчання для розв'язання прикладних задач аналізу даних з урахуванням принципів відтворюваності, етики та сучасних підходів MLOps.
Компетентності, формуванню яких сприяє дисципліна	ЗК-2. Здатність вчитися і оволодівати сучасними знаннями. ЗК-7. Вміння виявляти, ставити та вирішувати проблеми. СК-7. Здатність використовувати та впроваджувати нові технології, включаючи технології розумних, мобільних, зелених і безпечних обчислень, брати участь в модернізації та реконструкції комп'ютерних систем та мереж, різноманітних вбудованих і розподілених додатків, зокрема з метою підвищення їх ефективності.
Результати навчання	ПРН-3. Знати новітні технології в галузі комп'ютерної інженерії. ПРН-4. Знати та розуміти вплив технічних рішень в суспільному, економічному, соціальному і екологічному контексті. ПРН-8. Вміти системно мислити та застосовувати творчі здібності до формування нових ідей. ПРН-15. Вміти виконувати експериментальні дослідження за професійною тематикою. ПРН-20. Усвідомлювати необхідність навчання впродовж усього життя з метою поглиблення набутих та здобуття нових фахових знань, удосконалення креативного мислення.
Обсяг дисципліни	Загальний обсяг дисципліни: 3 кредити ЄКТС (90 годин). Для денної форми навчання: лекції – 14 годин, практичні заняття – 10 годин, лабораторні заняття – 10 годин, самостійна робота – 56 годин. Для заочної форми навчання: лекції – 4 години, практичні заняття – 4 години, лабораторні заняття – 4 години, самостійна робота –78 годин.
Форма підсумкового контролю	Залік
Терміни викладання дисципліни	Дисципліна викладається у 7-му семестрі.

Програма дисципліни

Тема 1.	**Ландшафт та методологія науки про дані.** Основні поняття та визначення. Історичні передумови та основні етапи розвитку галузі Data Science. Тенденції, що визначають майбутнє у сфері Data Science. Міждисциплінарний характер науки про дані (математика, статистика, обчислювальна техніка, розуміння конкретної галузі застосування). Робочий процес науки про дані: типові етапи (визначення проблеми, збирання даних, очищення даних, дослідницький аналіз даних EDA, проектування параметрів, побудова моделі, оцінка, розгортання, моніторинг). Ролі спеціалістів в галузі науки про дані: фахівець з обробки даних, інженер даних, інженер машинного навчання, бізнес-аналітик.

Тема 2.	**Збирання даних та бази даних.** Google Cloud, CloudSQL, Oracle, Microsoft SQL Server. API для збирання даних, методи веб-скрапінгу та потоки даних у реальному часі. Хмарні платформи даних. Основні постачальники хмарних послуг (AWS, Azure, GCP) та їх служби обробки даних (S3, BigQuery, Snowflake, Redshift для сучасних сховищ даних/озер).
Тема 3.	**Розширена попередня обробка даних та проектування параметрів.** Обробка пропущених значень. Стратегії підстановки (середнє значення, медіана, мода, регресійне, kNN). Виявлення та обробка викидів. Методи виявлення та обробки викидів (правило IQR, Z-оцінка, ізоляційні ліси). Методи категоріального кодування. Пряме кодування (One-hot encoding), кодування міток (label encoding), цільове кодування (target encoding), порядкове кодування (ordinal encoding), їх сфери використання. Масштабування параметрів. Стандартизація – Z-оцінка (Z-score), мін-макс нормалізація (Min-Max scaling). Принципи проектування параметрів. Створення нових ознак з існуючих; поліноміальні ознаки (polynomial features), взаємодії (interaction terms), агрегації (aggregations).
Тема 4.	**Дослідницький аналіз даних (EDA) та візуалізація.** Комплексний EDA. Методи одновимірного, двовимірного та багатовимірного аналізу. Статистичні зведення. Показники середнього/медіани; асиметрія (skewness), ексцес (kurtosis), квартилі (quartiles). Розширені інструменти візуалізації. Бібліотеки Matplotlib, Seaborn, Plotly. Ефективна комунікація здобутих відомостей в результаті EDA.
Тема 5.	**Оцінка та вибір моделі.** Метрики для оцінки продуктивності моделей машинного навчання у задачах класифікації. Точність (accuracy), прецизійність (precision), повнота (recall), F1-оцінка (F1-score), ROC-крива (ROC curve), AUC, матриця плутанини (confusion matrix). Метрики для оцінки продуктивності моделей машинного навчання у задачах регресії: MSE, RMSE, MAE, R-squared. Методи перехресної валідації: перехресна перевірка по K-згортці (K-fold cross-validation), стратифікована K-згортка Stratified K-fold, метод виключення по одному Leave-One-Out. Компроміс між "упередженістю" та дисперсією Bias-Variance trade-off – основна концепція ефективності моделі.
Тема 6.	**Введення в ансамблеве навчання.** Bagging (Random Forests – випадкові ліси), boosting (машини градієнтного бустингу – XGBoost, LightGBM). Причини ефективності ансамблевих методів машинного навчання: зменшення упередженості та дисперсії, покращення узагальнюючої здатності моделі машинного навчання. Практичні застосування ансамблевих методів машинного навчання.
Тема 7.	**Введення в Scikit-learn.** Практичне застосування найпоширеніших алгоритмів машинного навчання з використанням Scikit-learn (лінійна та логістична регресія, SVM, kNN, дерева рішень). ML-конвеєри (Model Pipelines). Створення надійних робочих процесів для попередньої обробки та моделювання.
Тема 8.	**Введення в обробку природної мови (NLP).** Основи текстових даних. Токенізація, стоп-слова, стемінг, лематизація. Представлення тексту: мішок слів (BoW), TF-IDF. Проста класифікація тексту. Застосування стандартних моделей машинного навчання до текстових даних.

Тема 9.	**Основи аналізу часових рядів.** Характеристики часових рядів даних: тренд, сезонність, циклічність, нерегулярність. Основні моделі прогнозування: рухоме середнє, експоненційне згладжування. Метрики оцінки часових рядів (RMSE, MAE, MASE та ін.).
Тема 10	**Введення в розгортання моделей та концепції MLOps.** Основні принципи MLOps: контроль версій моделей, моніторинг продуктивності моделей, стратегії повторного навчання (retraining). Етичні аспекти штучного інтелекту / науки про дані: упередженість (bias), справедливість (fairness), прозорість (transparency), конфіденційність (privacy).

Список рекомендованих джерел

1. VanderPlas J. Python Data Science Handbook, 2nd Edition. O'Reilly, 2022. 588 p.
2. Theobald O. Machine Learning For Absolute Beginners: A Plain English Introduction, 3rd edition, O'Reilly, 2021. 169 p.
3. Mueller J., Massaron L. Machine Learning for Dummies, 2nd edition. Wiley, 2021. 464 p.
4. Peter Wentworth, Jeffrey Elkner, Allen B. Downey, Chris Meyers. Learn Python the right way: how to think like a computer scientist. Ritza, 2022. 457 p.
5. Data Science Grade X, Version 1.0. Microsoft, 2022. 60 p. URL : https://cbseacademic.nic.in/web_material/codeingDS/classX_DS_Student_Handbook.pdf (дата звернення 20.08.2024).
6. Victor Tikhonov, Eduard Siemens, Yevhen Vasiliu, Valery Sitnikov, Abdullah Taher, Olena Tykhonova, Kateryna Shulakova and Serhii Tikhonov. “Context-Defined Model of Open Systems Interaction for IoT Cybersecurity Issues Study” // Proc. of 12-th Int. Conf. on Applied Innovation in IT, 2024/11/30, Vol. 12, Issue 2, pp.35-44. 2024, ISSN: 2199-8876. (Scopus).
7. Тихонов В.И., Тахер А., Тихонова Е.В. Застосування принципів багаторівневої граматики для захисту персональних даних // матеріали другої всеукраїнської НПК "Перспективні напрями захисту інформації" (Одеса, 3-7 вересня 2016 р.). О.: ОНАЗ ім. О.С.Попова, 2016. с. 77-80.
8. Practical Guide To Data Visualization: Part 1.Using Python, Seaborn, and Matplotlib. URL : https://towardsdatascience.com/practical-guide-to-data-visualization-83e375b0037 (дата звернення 20.08.2024).

Інформація про консультації

Згідно визначеного розкладу: ауд. 402 або онлайн за посиланням
https://us04web.zoom.us/j/3185149804?pwd=TmUybHZZYzBRK2dleUQrNVhPaG1wdz09

Загальна схема оцінювання

<table>
<tr><td rowspan="2">Сума балів за всі види навчальної діяльності</td><td rowspan="2">Шкала ЄКТС</td><td colspan="2">Оцінка за національною шкалою</td><td rowspan="2">Нарахування балів</td><td>Бали нараховуються таким чином:</td></tr>
<tr><td>для іспиту</td><td>для заліку</td><td rowspan="6">Оцінювання знань здобувачів вищої освіти здійснюється
за 100-бальною шкалою і становить: за поточну успішність (участь у практичних заняттях, виконання практичних завдань, лабораторних та контрольних робіт) та за результати заліку/екзамену)</td></tr>
<tr><td>90-100</td><td>A</td><td>Відмінно</td><td rowspan="5">зараховано</td></tr>
<tr><td>82-89</td><td>B</td><td rowspan="2">Добре</td></tr>
<tr><td>74-81</td><td>C</td></tr>
<tr><td>64-73</td><td>D</td><td rowspan="2">Задовільно</td></tr>
<tr><td>60-63</td><td>E</td></tr>
<tr><td>35-59</td><td>FX</td><td>Незадовільно з можливістю повторного складання</td><td>Не зараховано з можливістю повторного складання</td></tr>
<tr><td>0-34</td><td>F</td><td>Незадовільно з обов’язковим повторним вивченням дисципліни</td><td>Не зараховано з обов’язковим повторним вивченням дисципліни</td></tr>
</table>

Політика опанування дисципліни

Відвідування занять: відвідування здобувачами навчальних занять є обов’язковим, запізнення на заняття на 15 хвилин і більше не допускається. При проведенні занять в онлайн-режимі присутність здобувача зараховується у разі включення ним камери та/або мікрофона.

Умови зарахування пропущених занять: зарахування пропущених практичних/лабораторних занять здійснюється за умови виконання та захисту відповідних завдань. До заліку допускаються здобувачі, які виконали практичні та лабораторні завдання. Здобувач, який не з’явився на екзамен або не був допущений на момент його проведення, має право повторно його пройти у визначений викладачем термін.

Дотримання принципів академічної доброчесності: Підготовка усіх завдань, письмових робіт і т. ін., що виконуються в межах дисципліни, здійснюється здобувачем вищої освіти самостійно, на засадах академічної доброчесності. У разі порушення здобувачем принципів академічної доброчесності робота оцінюється незадовільно та має бути виконана повторно.