ЗАПОРІЗЬКА ПОЛІТЕХНІКА
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Інженерно-фізичний факультет
Кафедра фізичного матеріалознавства

СИЛАБУС

вибіркової навчальної дисципліни кафедрального каталогу

ОБРОБКА ТА АНАЛІЗ ДАНИХ В МАТЕРІАЛОЗНАВСТВІ

Обсяг освітнього компоненту: 6 кредитів ЄКТС / 180 годин

Освітня програма «Матеріалознавство та термічна обробка металів»
другого (магістерського) рівня вищої освіти
Спеціальність – G8 «Матеріалознавство»

ІНФОРМАЦІЯ ПРО ВИКЛАДАЧА

Ткач Дар'я Володимирівна
завідувач кафедри фізичного матеріалознавства,
кандидат технічних наук, доцент

Контактна інформація:
- e-mail: darynka.odarkivna@gmail.com;
- кафедра фізичного матеріалознавства, ауд. 159а, вул. Жуковського, 64, м. Запоріжжя, 69063.

Час і місце проведення консультацій:
згідно з поточним графіком консультацій; ауд. 159а, за потреби – із застосуванням засобів дистанційного зв'язку.

ОПИС КУРСУ

Дисципліна «Обробка та аналіз даних в матеріалознавстві» є вибірковим практично орієнтованим курсом для здобувачів другого (магістерського) рівня вищої освіти. Обсяг дисципліни становить 6 кредитів ЄКТС (180 годин), з них 32 години лекцій, 32 години практичних занять і 116 годин самостійної роботи. Форма підсумкового контролю – залік.

Курс поєднує базову статистичну обробку експериментальних матеріалознавчих даних із сучасними підходами Materials Data Science та машинного навчання. Навчання побудовано за логікою «матеріалознавча задача → дані → очищення і представлення → модель → перевірка → фізична інтерпретація → інженерний висновок». У фокусі – дані про хімічний склад, технологічні параметри, мікроструктуру, фазовий склад і властивості матеріалів; експериментальні та розрахункові набори даних; відкриті матеріалознавчі бази; регресія, класифікація, кластеризація, зниження розмірності, нейронні мережі, комп'ютерний зір, аналіз наукової літератури, інверсний дизайн та інтерпретований штучний інтелект.

Практична частина виконується переважно у Python (Google Colab/Jupyter) на підготовлених навчальних і відкритих матеріалознавчих наборах даних. Основна увага приділяється не обсягу програмного коду, а коректності підготовки даних, вибору методу, валідації моделі, відтворюваності аналізу та фізичній інтерпретації отриманих результатів.

МЕТА, КОМПЕТЕНТНОСТІ ТА РЕЗУЛЬТАТИ НАВЧАННЯ

Мета курсу – сформувати у здобувачів системне розуміння методів обробки, аналізу та інтерпретації матеріалознавчих даних, а також практичні навички застосування статистичних методів, Python та алгоритмів машинного навчання для розв'язування матеріалознавчих задач.

Компетентності, формування яких забезпечує дисципліна:

К1. Здатність формулювати матеріалознавчу задачу як задачу аналізу даних, визначати цільові змінні, інформативні ознаки, джерела даних та критерії якості результату.

К2. Здатність організовувати, структурувати, очищати й документувати експериментальні та розрахункові матеріалознавчі дані із забезпеченням їх відтворюваності.

К3. Здатність застосовувати методи описової статистики, візуалізації та дослідницького аналізу даних для виявлення закономірностей, аномалій і оцінювання варіабельності результатів.

К4. Здатність формувати фізично осмислені ознаки на основі хімічного складу, технологічних параметрів, характеристик структури та властивостей матеріалів.

К5. Здатність будувати, валідувати та порівнювати регресійні, класифікаційні й ансамблеві моделі машинного навчання, запобігаючи витоку даних і перенавчанню.

К6. Здатність застосовувати кластеризацію, методи зниження розмірності, нейронні мережі та комп'ютерний зір для аналізу матеріалознавчих даних і мікроструктурних зображень.

К7. Здатність працювати з відкритими матеріалознавчими базами даних і засобами автоматизованого вилучення інформації з наукових джерел із дотриманням принципів FAIR та відтворюваності.

К8. Здатність критично інтерпретувати результати моделей, оцінювати їх обмеження і невизначеність, застосовувати підходи інтерпретованого ШІ та формулювати матеріалознавчі й інженерні висновки.

Результати навчання:

РН1. Організовувати робоче середовище в Google Colab/Jupyter та використовувати базові можливості Python, NumPy, pandas і matplotlib для опрацювання матеріалознавчих даних.

РН2. Імпортувати, структурувати, очищати й документувати експериментальні та розрахункові дані, контролювати типи даних, одиниці вимірювання, пропущені значення, дублікати та аномалії.

РН3. Виконувати описову статистику й дослідницький аналіз даних, будувати та коректно інтерпретувати гістограми, boxplot, scatter plot, кореляційні матриці й матеріалознавчі діаграми.

РН4. Формувати матрицю ознак для задач машинного навчання, створювати матеріалознавчі дескриптори, виконувати кодування, масштабування та коректне розділення даних на навчальну і тестову вибірки.

РН5. Будувати та оцінювати регресійні й класифікаційні моделі за відповідними метриками, застосовувати крос-валідацію і базовий підбір гіперпараметрів та аналізувати помилки моделей.

РН6. Застосовувати методи кластеризації та зниження розмірності (PCA, t-SNE, UMAP) для пошуку груп, структурних подібностей і візуалізації багатовимірних матеріалознавчих просторів.

РН7. Застосовувати на базовому рівні нейронні мережі та методи комп'ютерного зору для прогнозування властивостей і аналізу мікроструктурних зображень.

РН8. Отримувати й перевіряти дані з відкритих матеріалознавчих баз та наукових публікацій, фіксувати походження даних і забезпечувати можливість відтворення аналізу.

РН9. Інтерпретувати результати моделей за допомогою feature importance та SHAP-підходів, оцінювати невизначеність і межі застосовності та відрізняти статистичну залежність від фізично обґрунтованого причинного висновку.

РН10. Виконувати наскрізний materials-data-проєкт: постановка задачі → підготовка даних → моделювання → валідація → інтерпретація → матеріалознавчий висновок.

ПЕРЕДУМОВИ ДЛЯ ВИВЧЕННЯ ДИСЦИПЛІНИ

Для успішного опанування дисципліни достатньо базових знань з матеріалознавства, фізики, математики та початкових навичок роботи з персональним комп'ютером. Бажаними є навички роботи з електронними таблицями. Попередній досвід програмування або машинного навчання не є обов'язковим: необхідні конструкції Python і базові алгоритми опановуються під час практичних занять.

ПЕРЕЛІК ТЕМ (ТЕМАТИЧНИЙ ПЛАН) ДИСЦИПЛІНИ

Таблиця 1 – Загальний тематичний план аудиторної роботи

Номер тижня	Теми лекцій, год.	Теми практичних робіт, год.
Змістовий модуль 1. Основи Materials Data Science та статистичний аналіз		
1	Вступ до матеріалознавчої інформатики. Еволюція наукових парадигм: спостереження, теорія, моделювання та data-driven science. Типи матеріалознавчих даних. Ланцюг «processing – structure – properties – performance». (2 год.)	Пр. № 1. Налаштування робочого середовища в Google Colab/Jupyter. Основи Python для дослідника: змінні, масиви, таблиці, імпорт CSV/XLSX; NumPy та pandas. (2 год.)
2	Статистичний огляд даних. Випадкові величини, типи шкал, середнє, медіана, мода, дисперсія, стандартне відхилення, квартилі, довірчі інтервали та похибки вимірювань у матеріалознавстві. (2 год.)	Пр. № 2. Дослідницький аналіз даних (EDA): описова статистика, гістограми, boxplot, scatter plot, кореляції; аналіз розподілу модуля Юнга та побудова спрощених карт Ешбі. (2 год.)
3	Джерела та архітектура матеріалознавчих даних. Експериментальні дані (SEM, EDS, XRD, EBSD, механічні випробування), симуляційні дані (DFT, MD, FEM), відкриті бази Materials Project і OQMD. Метадані, provenance та FAIR-принципи. (2 год.)	Пр. № 3. Збір і первинна організація матеріалознавчого датасету: структура таблиці, одиниці вимірювання, метадані, перевірка типів даних, уніфікація назв і форматів. (2 год.)
Змістовий модуль 2. Життєвий цикл даних і проєктування ознак		
4	Якість і очищення даних. Пропущені значення, дублікати, помилки введення, несумісні одиниці, викиди, data leakage та відтворюваність обробки. Статистичні й алгоритмічні підходи до аномалій. (2 год.)	Пр. № 4. Очищення даних у pandas: пропуски, дублікати, перетворення одиниць, IQR/z-score, виявлення аномалій; порівняння результатів до і після очищення. (2 год.)
5	Проєктування ознак (feature engineering). Матеріалознавчі дескриптори з хімічного складу, параметрів структури і процесу; кодування категорій, нормалізація і стандартизація; матриця ознак. (2 год.)	Пр. № 5. Підготовка датасету для прогнозування властивостей: формування Data Matrix, дескриптори складу, one-hot encoding, масштабування, train/test split. Початок наскрізного кейсу «склад сталі → властивість». (2 год.)
6	Візуальний і кореляційний аналіз перед моделюванням. Pair plots, heatmaps, мультиколінеарність, фізично осмислені трансформації та вибір ознак. Кореляція не дорівнює причинності. (2 год.)	Пр. № 6. Класичний аналіз датасету сталей: графіки «склад – властивість», кореляційна матриця, аналіз мультиколінеарності, формування базового набору ознак та baseline-моделі. (2 год.)
Змістовий модуль 3. Класичне машинне навчання в матеріалознавстві		
7	Регресійні моделі. Лінійна та поліноміальна регресія; MAE, RMSE, R^2; bias–variance trade-off; underfitting та overfitting. Матеріалознавчий зміст коефіцієнтів моделі. (2 год.)	Пр. № 7. Прогноз властивостей сталей за хімічним складом: лінійна/поліноміальна регресія, графіки predicted vs measured, residuals, MAE/RMSE/R^2, аналіз помилок. (2 год.)
8	Регуляризація та оцінювання моделей. Ridge, Lasso, Elastic Net, крос-валідація, підбір гіперпараметрів. Вступ до Gaussian Process Regression та оцінювання невизначеності. (2 год.)	Пр. № 8. Оптимізація регресійної моделі: k-fold cross-validation, GridSearchCV/RandomizedSearchCV, Ridge/Lasso/Elastic Net; порівняння стабільності та генералізації. (2 год.)
9	Класифікація. k-NN, дерева рішень, логістична регресія; confusion matrix, precision, recall, F1, ROC-AUC; дисбаланс класів. (2 год.)	Пр. № 9. Класифікація матеріалознавчих об'єктів або дефектів за табличними ознаками: train/test split, confusion matrix, F1, аналіз помилково класифікованих випадків. (2 год.)
10	Ансамблеві моделі. Random Forest, Gradient Boosting та XGBoost/HistGradientBoosting. Feature importance, стійкість до нелінійностей та ризики некоректного порівняння моделей. (2 год.)	Пр. № 10. Порівняння моделей для прогнозу властивостей матеріалу: baseline, Random Forest і gradient boosting; єдина схема крос-валідації, таблиця метрик, feature importance. (2 год.)
Змістовий модуль 4. Некероване навчання, нейронні мережі та комп'ютерний зір		
11	Кластеризація (unsupervised learning). k-means, DBSCAN, Gaussian Mixture Models; метрики відстані, масштабування, silhouette score та фізична інтерпретація кластерів. (2 год.)	Пр. № 11. Аналіз наноіндентування або локальних властивостей: кластеризація точок вимірювання для ідентифікації груп/фаз, порівняння k-means, DBSCAN і GMM. (2 год.)

Номер тижня	Теми лекцій, год.	Теми практичних робіт, год.
12	Зниження розмірності. PCA, t-SNE та UMAP: принципи, області коректного використання, втрати інформації та небезпека «красивих, але хибних» 2D-проєкцій. (2 год.)	Пр. № 12. Візуалізація багатовимірного простору матеріалів: PCA → t-SNE/UMAP, групування за класом матеріалу або властивістю, інтерпретація компонент та обмежень. (2 год.)
13	Штучні нейронні мережі (ANN/MLP). Перцептрон, шари, функції активації, backpropagation, регуляризація, validation set та рання зупинка. (2 год.)	Пр. № 13. Побудова простої MLP-моделі для прогнозування властивості матеріалу (наприклад, температури склування полімерів); криві навчання, validation loss, порівняння з класичними моделями. (2 год.)
14	Комп'ютерний зір у матеріалознавстві. Зображення як тензори, CNN, transfer learning, класифікація і сегментація мікроструктур; вимоги до розмітки та валідації. (2 год.)	Пр. № 14. Аналіз мікроструктурних зображень: підготовка набору, аугментація, transfer learning або сегментація; оцінювання результатів і зв'язок із кількісною металографією. (2 год.)
Змістовий модуль 5. Literature mining, інверсний дизайн, автономні лабораторії та інтерпретований ШІ		
15	NLP та Literature Mining у матеріалознавстві. Пошук і автоматизоване вилучення складів, режимів обробки, структурних характеристик і властивостей з наукових текстів; обмеження LLM та необхідність верифікації джерел. (2 год.)	Пр. № 15. Автоматизоване вилучення матеріалознавчих даних із тексту: регулярні вирази/простий NLP для складів і властивостей; формування структурованої таблиці з перевіркою першоджерел. (2 год.)
16	Генеративні моделі та інверсний дизайн. Active learning, Bayesian optimization, self-driving laboratories; інтерпретований ШІ (XAI), SHAP, відтворюваність і межі довіри до моделей. Підсумкова інтеграція курсу. (2 год.)	Пр. № 16. Підсумковий materials-data-проєкт: постановка задачі, EDA, preprocessing, модель, крос-валідація, XAI/SHAP, фізична інтерпретація, висновки та короткий захист. (2 год.)

САМОСТІЙНА РОБОТА

Самостійна робота становить 116 годин і спрямована на закріплення статистичних і матеріалознавчих основ, підготовку до практичних занять, виконання обчислювальних ноутбуків, роботу з науковими джерелами та відкритими базами даних, а також підготовку підсумкового materials-data-проєкту. Орієнтовний розподіл часу:

1. Python, Jupyter/Google Colab, NumPy і pandas для дослідницької роботи – 8 год.
2. Описова статистика, EDA, візуалізація та карти властивостей матеріалів – 8 год.
3. Джерела даних, Materials Project/OQMD, метадані, FAIR і відтворюваність – 8 год.
4. Очищення, контроль якості, аномалії та підготовка експериментальних даних – 8 год.
5. Feature engineering і матеріалознавчі дескриптори – 10 год.
6. Регресія, регуляризація, крос-валідація та оцінювання невизначеності – 10 год.
7. Класифікація, ансамблеві моделі та порівняння алгоритмів – 10 год.
8. Кластеризація, PCA, t-SNE та UMAP – 10 год.
9. Нейронні мережі та комп'ютерний зір для мікроструктур – 12 год.
10. NLP і literature mining для автоматизованого вилучення даних – 8 год.
11. Інверсний дизайн, active learning, XAI/SHAP і межі довіри до моделей – 10 год.
12. Підсумковий наскрізний materials-data-проєкт – 14 год.

РЕКОМЕНДОВАНІ ІНФОРМАЦІЙНІ ТА НАВЧАЛЬНО-МЕТОДИЧНІ ДЖЕРЕЛА

Для опрацювання курсу використовуються підготовлені викладачем конспекти лекцій, методичні матеріали до практичних робіт, документація програмних бібліотек і наведені нижче базові джерела.

Основні літературні джерела:

1. Agrawal A., Choudhary A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. APL Materials. 2016. Vol. 4. 053208. DOI: 10.1063/1.4946894.

2. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. Machine learning for molecular and materials science. Nature. 2018. Vol. 559. P. 547–555. DOI: 10.1038/s41586-018-0337-2.

3. Schmidt J., Marques M. R. G., Botti S., Marques M. A. L. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials. 2019. Vol. 5. Article 83. DOI: 10.1038/s41524-019-0221-0.

4. Ward L., Dunn A., Faghaninia A. et al. Matminer: An open source toolkit for materials data mining. Computational Materials Science. 2018. Vol. 152. P. 60–69. DOI: 10.1016/j.commatsci.2018.05.018.

5. Dunn A., Wang Q., Ganose A., Dopp D., Jain A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Computational Materials. 2020. Vol. 6. Article 138. DOI: 10.1038/s41524-020-00406-3.

6. Jain A., Ong S. P., Hautier G. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials. 2013. Vol. 1. 011002. DOI: 10.1063/1.4812323.

7. Saal J. E., Kirklin S., Aykol M., Meredig B., Wolverton C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). JOM. 2013. Vol. 65. P. 1501–1509. DOI: 10.1007/s11837-013-0755-4.

8. Wilkinson M. D., Dumontier M., Aalbersberg I. J. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016. Vol. 3. Article 160018. DOI: 10.1038/sdata.2016.18.

9. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

10. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.

11. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. 2nd ed. Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

12. Офіційна документація та навчальні ресурси: Python, Jupyter, Google Colab, NumPy, pandas, matplotlib, scikit-learn, matminer, Materials Project, OQMD, Matbench. Актуальні версії документації використовуються відповідно до календарного плану курсу.

ОЦІНЮВАННЯ

Поточний контроль передбачає оцінювання теоретичної підготовки, результатів виконання практичних робіт і підсумкового інтегрованого проєкту. Під час оцінювання враховуються коректність підготовки даних і вибору методу, відтворюваність аналізу, правильність використання метрик і валідації, якість візуалізації, фізична інтерпретація результатів та обґрунтованість матеріалознавчого висновку. Обсяг програмного коду сам по собі не є критерієм оцінювання.

Рекомендований розподіл балів: виконання та захист 16 практичних робіт – до 60 балів; два модульні контрольні заходи – до 20 балів; підсумковий materials-data-проєкт – до 20 балів. Загальна кількість – 100 балів.

Умови допуску до підсумкового контролю: виконання і захист обов’язкових практичних робіт та набір мінімальної кількості балів, установленої чинними нормативними документами університету. Вид підсумкового контролю – залік.

ПОЛІТИКИ КУРСУ

Організація навчання та комунікація. Основні навчальні матеріали, набори даних, завдання для самостійної роботи, критерії оцінювання та терміни виконання розміщуються в системі Moodle. Консультації проводяться відповідно до графіка консультацій викладача, а за необхідності – із застосуванням засобів дистанційного зв’язку.

Практичні та проєктні роботи. Кожна робота має містити вихідні дані або посилання на їх джерело, відтворюваний код, коротке пояснення послідовності обробки, метрики та/або графіки і матеріалознавчий висновок. Здобувач повинен уміти пояснити логіку виконаних операцій і отриманих результатів.

Академічна доброчесність. Код, текст, рисунки, дані та моделі інших авторів мають супроводжуватися коректними посиланнями. Не допускаються академічний плагіат, фабрикація або фальсифікація даних, списування, виконання оцінюваної роботи іншою особою чи подання чужих ноутбуків як власних.

Використання генеративного штучного інтелекту. Генеративні інструменти можуть використовуватися для пояснення синтаксису, пошуку помилок, підготовки чернеток коду або структурування тексту, якщо це не суперечить умовам конкретного завдання. У разі використання ШІ здобувач самостійно перевіряє достовірність інформації, працездатність коду й коректність висновків та дотримується вимог університету щодо зазначення такого використання.

Дедлайни та повторне виконання робіт. Терміни виконання навчальних завдань повідомляються здобувачам заздалегідь і зазначаються в Moodle. За наявності обґрунтованих обставин термін може бути змінений за погодженням із викладачем. Роботи, за які отримано результат нижчий від мінімально встановленого рівня, можуть бути доопрацьовані та повторно подані відповідно до встановленого порядку.

Оскарження результатів оцінювання. Здобувач має право отримати від викладача пояснення щодо результатів оцінювання та оскаржити результати контрольного заходу відповідно до нормативних документів НУ «Запорізька політехніка».

Академічна мобільність. Результати навчання, отримані в межах програм академічної мобільності або сертифікованих курсів із data science та machine learning, можуть бути визнані відповідно до встановленого в університеті порядку.

ТЕХНІЧНІ ВИМОГИ ДЛЯ РОБОТИ НА КУРСІ

Для роботи на курсі потрібні доступ до Moodle та Google Colab або локального середовища Jupyter, персональний комп’ютер із сучасним веббраузером і стабільним доступом до мережі. Основне програмне середовище – Python 3 із бібліотеками NumPy, pandas, matplotlib, scikit-learn; для окремих тем можуть використовуватися seaborn, scipy, umap-learn, shap, TensorFlow/Keras або PyTorch, matminer та інші відкриті бібліотеки. Спеціалізоване платне програмне забезпечення для виконання обов’язкових завдань не потрібне.

Навчальні набори даних добираються з відкритих і перевірюваних джерел: Materials Project, OQMD, Matbench, наукових статей із доступними supplementary data, а також зі знеособлених кафедральних експериментальних наборів. Для кожного набору даних має бути зафіксовано його походження, одиниці вимірювання та основні умови отримання.