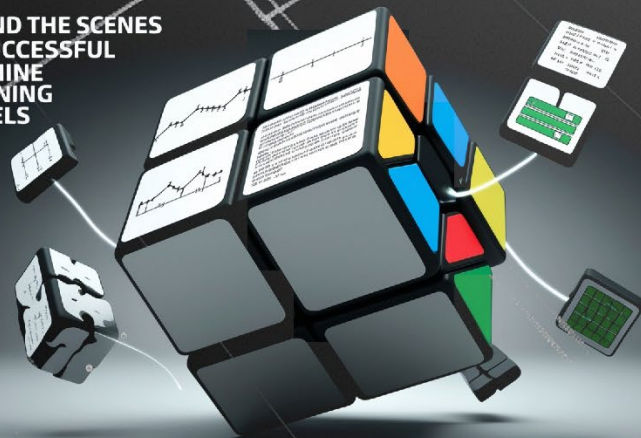


Марія Талах

ПРИКЛАДНЕ МАШИННЕ НАВЧАННЯ

Навчально-методичний посібник

BEHIND THE SCENES
OF SUCCESSFUL
MACHINE
LEARNING
MODELS



КАФЕДРА КОМП'ЮТЕРНИХ НАУК



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧЕРНІВЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЮРІЯ ФЕДЬКОВИЧА

М.В. Талах

ПРИКЛАДНЕ МАШИННЕ НАВЧАННЯ

Навчально-методичний посібник

Механодрук

2025

УДК 004.85

T-16

Рекомендовано до друку вченою радою навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького національного університету імені Юрія Федьковича (протокол № 17 від 18 грудня 2025 р.)

Рецензенти:

Рецензенти: Баловсяк С.В., доктор технічних наук, доцент, доцент кафедри комп'ютерних систем та мереж

Виклюк Я.І., доктор технічних наук, професор, професор кафедри систем штучного інтелекту Національного університету «Львівська політехніка»

Талах М.В.

T-16 Прикладне машинне навчання : навчальний посібник / М.В. Талах – Чернівці: Технодрук, 2025. – 408 с.

Навчальний посібник присвячений комплексному вивченню прикладних аспектів машинного навчання, охоплюючи методологію CRISP-ML(Q), запобігання витоку даних, інженерію ознак, роботу з незбалансованими даними, налаштування гіперпараметрів, інтерпретацію моделей та використання сучасних фреймворків для обробки мультимодальних даних (текст, зображення, послідовності, аудіо), а також автоматизоване машинне навчання й інтеграцію з великими мовними моделями. Він систематизує сучасні підходи до розробки надійних ML-систем корпоративного рівня, включаючи забезпечення якості даних, валідацію моделей і підвищення інтерпретованості, та призначений для магістрів спеціальності 122 "Комп'ютерні науки" (освітня програма «Інтелектуальний аналіз даних в комп'ютерних інформаційних системах») і фахівців, які прагнуть поглибити знання в прикладному машинному навчанні.

УДК 004.85

©Чернів. нац. ун-т, 2025

©ПВКФ «Технодрук», 2025

©Талах М.В., 2025

ЗМІСТ

ВСТУП	9
<u>РОЗДІЛ 1. CRISP-ML(Q): МЕТОДОЛОГІЯ РОЗРОБКИ ML-ПРОЄКТІВ З ЗАБЕЗПЕЧЕННЯМ ЯКОСТІ</u>	10
1.1. СТРУКТУРА МЕТОДОЛОГІЇ CRISP-ML(Q)	11
1.2. ФАЗИ CRISP-ML(Q)	13
1.2.1. РОЗУМІННЯ БІЗНЕСУ ТА ДАНИХ	13
1.2.2. ІНЖЕНЕРІЯ ДАНИХ (ПІДГОТОВКА ДАНИХ)	14
1.2.3. РОЗРОБКА МОДЕЛІ МАШИННОГО НАВЧАННЯ	15
1.2.4. ОЦІНКА МОДЕЛІ МАШИННОГО НАВЧАННЯ	17
1.2.5. РОЗГОРТАННЯ	18
1.2.6. МОНІТОРИНГ ТА ОБСЛУГОВУВАННЯ	19
1.3. ПЕРЕВАГИ ВИКОРИСТАННЯ CRISP-ML(Q)	20
1.4. ІНТЕГРАЦІЯ CRISP-ML(Q) З MLOPS	22
1.5. ПРАКТИЧНЕ ЗАСТОСУВАННЯ CRISP-ML(Q)	22
1.5.1. ІНСТРУМЕНТИ ТА ПЛАТФОРМИ ДЛЯ ПІДТРИМКИ CRISP-ML(Q)	22
1.5.2. ВПРОВАДЖЕННЯ CRISP-ML(Q) В ОРГАНІЗАЦІЇ	23
1.6. ВИКЛИКИ ТА ОБМЕЖЕННЯ CRISP-ML(Q)	24
1.7. МАЙБУТНЄ МЕТОДОЛОГІЙ ML-РОЗРОБКИ	25
1.8. CRISP-ML(Q) ЯК КОНЦЕПТУАЛЬНИЙ ФУНДАМЕНТ СУЧАСНОЇ ML-ПРАКТИКИ	26
<u>РОЗДІЛ 2. ВИТІК ДАНИХ: КОНЦЕПЦІЯ, ШЛЯХИ ВІЯВЛЕННЯ ТА ПОДОЛАННЯ</u>	29
2.1. РОЗУМІННЯ ВИТОКУ ДАНИХ	30
2.2. ЧОМУ ВІДБУВАЄТЬСЯ ВИТІК ДАНИХ?	32
2.2.1. ОСНОВНІ ДЖЕРЕЛА ВИТОКУ ДАНИХ	32
2.2.2. ВИТІК ДАНИХ ПІД ЧАС ПРЕПРОЦЕСИНГУ	35
2.3. ДЕТЕКЦІЯ ВИТОКУ ДАНИХ	35
2.3.1. ОСНОВНІ СЦЕНАРІЇ ВІЯВЛЕННЯ ВИТОКУ ДАНИХ	36
2.3.2. МЕТОДИ ВІЯВЛЕННЯ ВИТОКУ ДАНИХ	37
2.4. ЯК МІНІМІЗУВАТИ ВИТІК ДАНИХ?	38
2.4.1. ІДЕЯ 1: РЕТЕЛЬНИЙ ВИБІР ОЗНАК	38

2.4.2. Ідея 2: Створення окремого валідаційного набору	39
2.4.3. Ідея 3: Правильна послідовність препроцесингу даних	39
2.4.4. Ідея 4: Особливості роботи з часовими рядами	40
2.4.5. Ідея 5: Правильне використання перехресної валідації	41
2.5. Інші методи запобігання витоку даних	42
2.6. Витік даних у складних сценаріях	44
2.7. Реальні приклади витоку даних у проєктах машинного навчання	45
2.7.1. Витік даних у навчальних даних	45
2.7.2. Витік через проксі-змінні	46
2.7.3. Вплив сторонніх наборів даних	46
2.8. Висновки та рекомендації	47
2.8.1. Підсумок ключових стратегій	47
2.8.2. Інтеграція з методологією CRISP-ML(Q)	48
2.8.3. Заключні рекомендації	49

РОЗДІЛ 3. РОБОТА З ОЗНАКАМИ (FEATURE ENGINEERING) ДЛЯ МАШИННОГО НАВЧАННЯ

3.1. FEATURE ENGINEERING	52
3.1.1. Цілі та переваги FEATURE ENGINEERING	53
3.1.2. Підходи до FEATURE ENGINEERING	54
3.2. Екстракція ознак	58
3.2.1. Сутність екстракції ознак	58
3.2.2. Теоретичні основи та підходи до екстракції ознак	59
3.2.3. Екстракція ознак з текстових даних	61
3.2.4. Екстракція ознак з зображень	70
3.2.5. Екстракція ознак з послідовних даних	84
3.2.6. Екстракція ознак із аудіоданих	102
3.3. Відбір ознак	110
3.3.1. Категоризація методів відбору ознак	111

3.3.2 ФІЛЬТРАЦІЙНІ МЕТОДИ (FILTER METHODS)	111
3.3.3 ОБГОРТКОВІ МЕТОДИ (WRAPPER METHODS)	115
3.3.4 ВБУДОВАНІ МЕТОДИ (EMBEDDED METHODS)	118
3.3.5 РЕГУЛЯРИЗАЦІЙНІ МЕТОДИ	118
3.4 СУЧАСНІ ПІДХОДИ ДО ВІДБОРУ ОЗНАК	121
3.4.1 ГІБРИДНІ МЕТОДИ ВІДБОРУ ОЗНАК	121
3.4.2 АНСАМБЛЕВІ МЕТОДИ ВІДБОРУ ОЗНАК	121
3.5 ВІДБІР ОЗНАК ЗАЛЕЖНО ВІД ТИПІВ ДАНИХ	122
3.5.1 ВІДБІР ОЗНАК ДЛЯ РІЗНИХ КОМБІНАЦІЙ ТИПІВ ДАНИХ	122
3.5.2 УНІВЕРСАЛЬНІ МЕТОДИ ВІДБОРУ ОЗНАК	125
3.5.3. ПРАКТИЧНІ РЕКОМЕНДАЦІЇ ЩОДО ВІДБОРУ ОЗНАК	127
3.6 БІБЛІОТЕКИ ДЛЯ ВІДБОРУ ОЗНАК У PYTHON	129
3.6.1. SCIKIT-LEARN	130
3.6.2. ITMO FEATURE SELECTION LIBRARY	134
3.6.3. ARIZONA STATE UNIVERSITY FEATURE SELECTION LIBRARY	134
3.6.4. FEATURE ENGINEERING AND SELECTION (FES)	134
3.6.5. MLR3	134
3.6.6. ІНШІ ВАЖЛИВІ БІБЛІОТЕКИ	135
3.7. Підсумки та рекомендації	136

РОЗДІЛ 4. РОБОТА З НЕЗБАЛАНСОВАНИМИ ДАНИМИ **138**

4.1. ПРОБЛЕМА ДИСБАЛАНСУ КЛАСІВ У МАШИННОМУ НАВЧАННІ	140
4.2. МЕТОДИ ПОВТОРНОЇ ВИБІРКИ	143
4.2.1. БАЗОВІ МЕТОДИ ПОВТОРНОЇ ВИБІРКИ	143
4.2.2. СУЧАСНІ МЕТОДИ ОВЕРСЕМПЛІНГУ	144
4.2.3. СУЧАСНІ МЕТОДИ АНДЕРСЕМПЛІНГУ	145
4.2.4. ГІБРИДНІ МЕТОДИ	146
4.3. АНСАМБЛЕВІ МЕТОДИ ДЛЯ НЕЗБАЛАНСОВАНИХ ДАНИХ	147
4.3.1. BALANCED RANDOM FOREST	147
4.3.2. EASYENSEMBLE	149
4.3.3. XGBOOST ІЗ ВАГАМИ КЛАСІВ	150

4.4. МЕТРИКИ ОЦІНКИ ДЛЯ НЕЗБАЛАНСОВАНИХ ДАНИХ	152
4.4.1. F1-SCORE	152
4.4.2. G-MEAN	153
4.4.3. AUROC	153
4.4.4. AUPRC	153
4.5. АДАПТИВНІ ТА ГІБРИДНІ ПІДХОДИ	155
4.5.1. ПАЙПЛАЙНИ З IMBALANCED-LEARN	155
4.5.2. АВТОМАТИЗАЦІЯ ВИБОРУ СТРАТЕГІЙ ІЗ GRIDSEARCHCV	156
4.5.3. ГІБРИДНІ МЕТОДИ ЯК АДАПТИВНЕ РІШЕННЯ	157
4.6. ПРАКТИЧНІ РЕКОМЕНДАЦІЇ ТА ІТЕРАТИВНІСТЬ У CRISP-ML(Q)	158
4.7. НАВЧАННЯ З УРАХУВАННЯМ ВАРТОСТІ ПОМИЛОК (COST-SENSITIVE LEARNING)	161
4.7.1. МАТРИЦЯ ВИТРАТ І ЇЇ РОЛЬ	162
4.7.2. МЕТОДИ НАВЧАННЯ З УРАХУВАННЯМ ВАРТОСТІ ПОМИЛОК	163
4.7.3. ПЕРЕВАГИ ТА ОБМЕЖЕННЯ	169
4.7.4. ІТЕРАТИВНІСТЬ У CRISP-ML(Q)	170
4.8. ПІДСУМКИ ТА ПРАКТИЧНІ РЕКОМЕНДАЦІЇ ЩОДО РОБОТИ З НЕЗБАЛАНСОВАНИМИ ДАНИМИ	171

РОЗДІЛ 5. НАЛАШТУВАННЯ МОДЕЛІ, ТЕХНІКИ ПІДБОРУ ОПТИМАЛЬНИХ ПАРАМЕТРІВ

5.1. ПЕРЕХРЕСНА ПЕРЕВІРКА	176
5.1.1. К-БЛОКОВА ПЕРЕХРЕСНА ПЕРЕВІРКА	177
5.1.2. СТРАТИФІКОВАНА К-БЛОКОВА ПЕРЕХРЕСНА ПЕРЕВІРКА	180
5.1.3. ПЕРЕХРЕСНА ПЕРЕВІРКА З ВИКЛЮЧЕННЯМ ПО ОДНОМУ (LEAVE-ONE-OUT)	181
5.1.4. ІНШІ СТРАТЕГІЇ ПЕРЕХРЕСНОЇ ПЕРЕВІРКИ	182
5.2. СІТКОВИЙ ПОШУК ТА АЛЬТЕРНАТИВНІ МЕТОДИ ПІДБОРУ ПАРАМЕТРІВ	185
5.2.1. АЛЬТЕРНАТИВНІ МЕТОДИ ПІДБОРУ ПАРАМЕТРІВ	191
5.2.2. ВКЛАДЕНА ПЕРЕХРЕСНА ПЕРЕВІРКА	194

5.3. CRISP-ML(Q)-ОРІЄНТОВАНИЙ ЧЕК-ЛИСТ НАЛАШТУВАННЯ МОДЕЛЕЙ: ВАЛІДАЦІЯ, ПІДБІР І АВТОМАТИЗАЦІЯ	195
--	-----

РОЗДІЛ 6. ПІДХОДИ ДО ІНТЕРПРЕТАЦІЇ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

6.1. ПОНЯТТЯ ІНТЕРПРЕТОВАНOSTІ	199
6.2. ОЦІНКА ІНТЕРПРЕТОВАНOSTІ	201
6.3. ОСНОВНІ МЕТОДИ ІНТЕРПРЕТАЦІЇ	201
6.3.1. LIME (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS)	202
6.3.2. SHAP (SHAPLEY ADDITIVE EXPLANATIONS)	206
6.3.3. ВІЗУАЛІЗАЦІЯ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ	210
6.4. ПОРІВНЯННЯ МЕТОДІВ ІНТЕРПРЕТАЦІЇ	213
6.5. ПРАКТИЧНИЙ ПРИКЛАД: АНАЛІЗ ДАНИХ ПРО ДІАБЕТ	214
6.6. ІНТЕРПРЕТОВАНІСТЬ МОДЕЛЕЙ У CRISP-ML(Q): ПІДСУМКИ ТА ПРИКЛАДНІ РЕКОМЕНДАЦІЇ	216

РОЗДІЛ 7. ВИСОКОРІВНЕВІ АБСТРАКЦІЇ ТА ФРЕЙМВОРКИ МАШИННОГО НАВЧАННЯ

7.1. ІНТЕЛЕКТУАЛЬНІ АБСТРАКЦІЇ ДЛЯ АНАЛІЗУ ДАНИХ	221
7.1.1. PANDASAI ТА АВТОМАТИЗАЦІЯ АНАЛІЗУ ДАНИХ	221
7.1.2. ІНТЕГРАЦІЯ З LLM	309
7.1.3. ПОРІВНЯННЯ З ТРАДИЦІЙНИМИ МЕТОДАМИ EDA	313
7.2. АБСТРАКЦІЇ МОДЕЛЮВАННЯ ТА НАВЧАННЯ	317
7.2.1. ОСНОВНІ КОНЦЕПЦІЇ АБСТРАКЦІЙ МОДЕЛЮВАННЯ ТА НАВЧАННЯ	317
7.2.2. PYTORCH IGNITE: CREATE_SUPERVISED_TRAINER	318
7.2.3. PYTORCH LIGHTNING	319
7.2.4. KERAS ТА TENSORFLOW EXTENDED	322
7.2.5. FASTAI	324
7.2.6. ПОРІВНЯЛЬНИЙ АНАЛІЗ ПРОДУКТИВНОСТІ ТА ЗРУЧНОСТІ	326
7.3. АБСТРАКЦІЇ КОНВЕЄРІВ ОБРОБКИ ДАНИХ	329
7.3.1. MLFLOW	330

7.3.2. DATA VERSION CONTROL (DVC)	332
7.3.3. APACHE AIRFLOW для ML-конвексів	334
7.3.4. ZENML	337
7.4. АБСТРАКЦІЇ РОЗГОРТАННЯ ТА МОНІТОРИНГУ	342
7.4.1. TENSORFLOW SAVEDMODEL	343
7.4.2. PYTORCH JIT	343
7.4.3. BENTOML	344
7.4.4. SELDON CORE	347
7.4.5. KSERVE (РАНІШЕ KFSERVING)	350
7.4.6. СИСТЕМИ МОНІТОРИНГУ ТА А/В ТЕСТУВАННЯ	354
7.4.7. ПОРІВНЯННЯ АБСТРАКЦІЙ РОЗГОРТАННЯ ТА МОНІТОРИНГУ ML-СИСТЕМ	358
7.5. ВИСОКОРІВНЕВІ БІБЛІОТЕКИ ДЛЯ СПЕЦИФІЧНИХ ЗАВДАНЬ	360
7.5.1. HUGGINGFACE TRANSFORMERS для NLP	361
7.5.2. MONAI для МЕДИЧНИХ ЗОБРАЖЕНЬ	365
7.5.3. ТІММ для КОМП'ЮТЕРНОГО ЗОРУ	370
7.5.4. GNN-БІБЛІОТЕКИ для ГРАФОВИХ ЗАДАЧ	375
7.5.5. ПОРІВНЯЛЬНИЙ АНАЛІЗ ВИСОКОРІВНЕВИХ БІБЛІОТЕК для СПЕЦИФІЧНИХ ЗАВДАНЬ	383
7.6. AUTOML: АВТОМАТИЗАЦІЯ ПРОЦЕСІВ МАШИННОГО НАВЧАННЯ	386
7.6.1. ЗАГАЛЬНИЙ ПРИНЦИП РОБОТИ AUTOML	386
7.6.2. ЗАСТОСУВАННЯ AUTOML на ЕТАПАХ ЖИТТЄВОГО ЦИКЛУ ML	388
7.7. ПРАКТИЧНІ ОРІЄНТИРИ за ФОРМАТАМИ ДАНИХ і ЗАДАЧАМИ	400
ВИСНОВКИ	404
ЛІТЕРАТУРА	405