Теоретическая часть

Методи штучного інтелекту включають методи, які дозволяють комп'ютерам імітувати людську поведінку, даючи їм змогу навчатися, приймати рішення, розпізнавати закономірності та вирішувати складні проблеми так, як це робить людський інтелект.

Машинне навчання – це підгрупа методів штучного інтелекту, що використовує вдосконалені алгоритми для виявлення закономірностей у великих наборах даних, дозволяючи машинам навчатися та адаптуватися. Таким чином у вузькому сенсі методи машинного навчання можна визначити як алгоритми та моделі, які використовуються для обробки та аналізу великих обсягів даних. Вони дозволяють комп'ютерним системам витягувати знання з даних і використовувати їх для прийняття рішень і формування прогнозів.

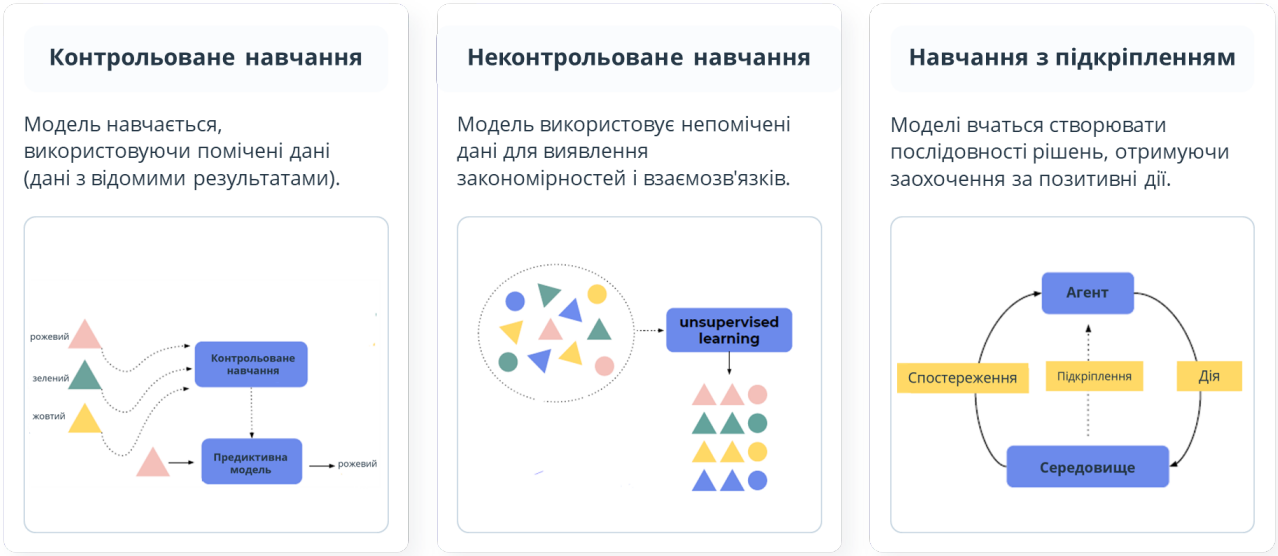
Іншими словами машинне навчання – це підхід до (1) вивчення (2) складних закономірностей на основі (3) наявних даних і використання цих закономірностей для (4) передбачення невідомих даних.

Наприклад, для задачі прогнозування врожайності пшениці відповідні блоки будуть спрямовані на вирішення наступних задач:

1. Алгоритм — встановлює зв'язки або вплив різних факторів, таких як кількість опадів, температура, використання добрив і технології обробки ґрунту, тощо на врожайність пшениці
2. Дані — історичні спостереження за врожайністю пшениці, погодні умови кожного сезону, використання добрив, процедури обробки ґрунту, попередні культури і інші агротехнічні заходи
3. Модель — сталий вплив на врожайність певних діапазонів температур та кількості опадів, впливи пов’язані із взаємодією різними фаторів, як-то видами добрив та типами ґрунту, попередні культури, тощо
4. Прогнози — очікувана врожайність в наступному сезоні на певному грунті, якщо погодні умови та агротехнічні заходи будуть подібними до одного з попередніх сезонів.

Типи машинного навчання

Методи машинного навчання можна класифікувати на три основні категорії: навчання із учителем, навчання без учителя та навчання з підкріпленням. Навчання із учителем передбачає використання пар “вхідні дані – вихідні дані”, де модель навчається на основі розмічених даних, що дозволяє їй прогнозувати результати для нових спостережень. Навчання без учителя, на відміну від цього, працює з не розміченими даними, де алгоритми самостійно виявляють закономірності та структури в даних, наприклад, через кластеризацію або зменшення розмірності. Нарешті, навчання з підкріпленням зосереджене на навчанні агента через взаємодію з середовищем; агент виконує дії в певних станах і отримує винагороди або штрафи, що стимулює його до максимізації загальної нагороди через оптимізацію своїх дій.



Категорії машинного навчання

Навчання без учителя – це метод, при якому система навчається на основі не розмічених даних. Мета полягає у виявленні закономірностей та структури в даних без будь-яких попередньо заданих умов. Приклади таких методів включають кластеризацію та зменшення розмірності. Кластеризація – це метод, коли модель розділяє дані на групи (кластери) на основі їхньої схожості. Основними алгоритмами, що використовуються для кластеризації, є k-середніх і DBSCAN. Зниження розмірності, в свою чергу, дозволяє зменшити кількість ознак у даних, зберігаючи при цьому основну інформацію. Методи навчання без учителя можуть застосовуватися в різних сферах, таких як аналіз даних, пошук аномалій, стиснення даних та інші.

Методи навчання без учителя

Навчання з учителем – це метод, при якому система навчається на основі пар «вхідні дані – вихідні дані». Це означає, що модель отримує вхідні дані разом із правильними відповідями, що дозволяє їй навчитися передбачати правильні результати для нових вхідних даних. Наприклад, у випадку прогнозування урожайності сільськогосподарських культур, система може використовувати дані про погоду, тип ґрунту та рівень добрив, щоб навчитися передбачати ймовірний урожай. Основні етапи процесу навчання з учителем включають підготовку даних, вибір моделі, навчання, оцінку моделі та використання для прогнозів. Якість моделі залежить від якості даних, на яких вона навчається, а також від правильного вибору моделі і параметрів.

Навчання з учителем в свою чергу ділиться на дві підгрупи: моделі (або алгоритми) регресії та моделі (або алгоритми) класифікації.

Регресія – це, по суті, спосіб передбачення майбутніх числових значень на основі минулих ознак. Вона допомагає прогнозувати такі показники, як продажі, ціни, врожайність, робити рекомендації. Основні методи регресії наведено на рис:

Базові методи регресії

Методи класифікації на відміну допомагають передбачити, до якої категорії чи групи належить кожне спостереження. Ці методи можуть допомогти визначити, чи є певні типи хвороб у рослин чи тварин, чи електронний лист спамом, або передбачити, чи має банк схвалити кредитну заявку.

Базові методи класифікації

Для складних задач у машинному навчанні часто є недостатньо лише однієї моделі, оскільки одна модель може не повністю впоратися з усіма аспектами або нюансами

досліджуваної проблеми. У таких випадках використовують ансамблеві методи, які об'єднують кілька моделей для покращення загальної точності та надійності прогнозів. Основні типи ансамблевих методів включають методи узгодження (bagging), такі як Random Forest, які створюють численні моделі з випадкових підвибірок навчальних даних і комбінують їхні результати для зменшення варіативності та підвищення стабільності. Іншим підходом є методи підсилювання (boosting), як-от AdaBoost або Gradient Boosting, які навчають моделі послідовно, зосереджуючи увагу на прикладах, які були класифіковані неправильно попередніми моделями, щоб покращити загальні результати. Ці ансамблеві техніки часто призводять до значного збільшення продуктивності ансамбля моделей, оскільки окремі моделі мають свої сильні сторони та дозволяють компенсувати слабкості інших.

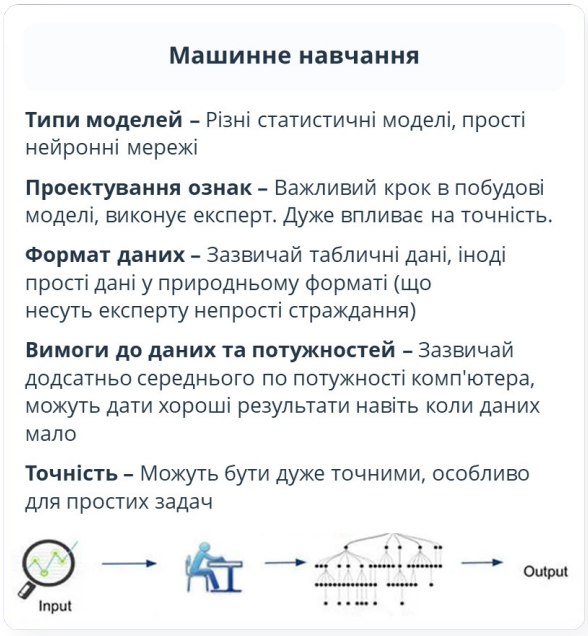
Ансамблеві методи

Навчання з підкріпленням — цей тип навчання складається з декількох етапів: визначення задачі, визначення станів і дій, функції нагороди, навчання моделі, оцінки моделі та використання. Основними методами навчання з підкріпленням є Q-навчання та методи глибокого навчання, такі як Deep Q-Networks (DQN) і Policy

Gradient. Q-навчання розраховує функцію Q(s,a), яка визначає очікувану винагороду за виконання певних дії за різних станах

Окремою галуззю машинного навчання є *Глибоке навчання.* Воно базується на використанні багатошарових нейронних мереж для автоматичного навчання і виявлення складних патернів у великих обсягах даних. Перехід від класичного машинного навчення до глибокого навчання зазвичай відбувається, коли розмір і складність даних перевищують можливості традиційних алгоритмів, а також коли є необхідність працювати з неструктурованими даними, такими як зображення, аудіо або текст. Основна відмінність глибокого навчання від класичних підходів полягає в здатності нейронних мереж автоматично витягувати ознаки з сирих даних без ручного відбору.

Глибокі моделі здатні вчитися на різних рівнях абстракції; вони можуть виявляти прості ознаки на нижчих шарах, а складніші патерни на верхніх. Глибоке навчання застосовується в широкому спектрі задач, таких як комп'ютерний зір, обробка природної мови та генерація контенту, вимагаючи значних обчислювальних потужностей і великих обсягів даних для навчання.



Основи управління даними в системах Штучного Інтелекту

Управління даними в системах штучного інтелекту є критично важливим етапом, оскільки якість та релевантність даних прямо впливають на ефективність і точність моделей штучного інтелекту. Основи управління даними включають кілька ключових етапів:

1. *Збір даних* – цей етап передбачає збирання даних з різноманітних джерел, які можуть включати бази даних, API, анкети, датчики та інше. Якість зібраних даних визначає успішність подальших етапів.

У контексті сільськогосподарських технологій збір даних може включати в себе такі методи та джерела, як Супутникові зображення, Дані від сенсорів на сільськогосподарській техніці, Метеодані, Текстові дані, Табличні дані

2. *Розмітка даних* – на цьому етапі дані отримують мітки, які описують їхній зміст. Це може бути виконане вручну експертами або автоматизовано за допомогою алгоритмів, залежно від задачі. Правильне маркування є критично важливим для навчання моделей, оскільки воно дозволяє системам навчатися на основі зрозумілих прикладів. З точки зору моделі машинного навчання Розмітка даних - це процес ідентифікації у вхідних даних одного чи кількох цільових признаків, значення яких має прогнозувати модель. Це дозволяє моделі МН вчитися на даних і розпізнавати закономірності. Для керуючих моделей навчання мітки є важливими, оскільки вони діють як прямі індикатори бажаного виходу.

3. *Розвідувальний аналіз даних* (Exploratory Data Analysis, EDA) – цей етап включає використання статистичних і візуалізаційних методів для дослідження зібраних даних. EDA допомагає виявити закономірності, аномалії та інші особливості даних, що можуть вплинути на подальший аналіз і розробку моделей.

Отримані в результаті узагальнення визначають наступні кроки підготовки даних та моделювання, забезпечуючи високу якість та глибоке розуміння даних, які подаються на вхід моделям машинного навчання.

4. Підготовка даних – на цьому етапі дані очищуються і трансформуються для подальшої обробки. Це може включати видалення пропущених значень, нормалізацію, кодування категоріальних змінних і масштабування числових даних. Якісна підготовка даних забезпечує, що моделі отримують коректну і чітко структуровану інформацію, що підвищує ймовірність успішного навчання.

Попередня обробка є важливою для перетворення сирої інформації в чистий, організований формат, який підходить для побудови надійних моделей машинного навчання. Це включає не лише очищення і перетворення даних, але також зменшення варіативності і доповнення їх для покращення результатів навчання моделі. Типові техніки попередньої обробки даних наведені на рис.

Очищення даних

Мета: Видалення неточностей та виправлення невідпові дностей в наборі даних.

Техніки: Робота з пропущеними значеннями, видалення дублікатів, виправлення непотрібних записів.

Трансформація даних

Мета: Стандартизація та нормалізація даних для забезпечення їхньої однорідності в усьому наборі даних.

Техніки: Масштабування ознак, кодування категоріальних змінних, застосування трансформацій для нормалізації розподілу даних.

Зменшення та розширення даних

Мета: Керування розміром та різноманітністю набору даних для забезпечення балансу між обчислювальною ефективністю та точністю моделі.

Техніки: Зменшення обсягу даних: Використання методів зменшення розмірності, таких як метод головних компонент або методи відбору ознак, для збереження лише найінформативніших ознак, зменшуючи складність моделі та час її тренування.

Розширеня даних: Штучне розширення тренувального набору даних шляхом створення реалістичних модифікацій та варіацій існуючих даних (наприклад, за допомогою технік, таких як SMOTE для табличних даних або трансформації зображень для візуальних даних), що допомагає підвищити загальну здатність моделі до узагальнення.

Перевірка даних

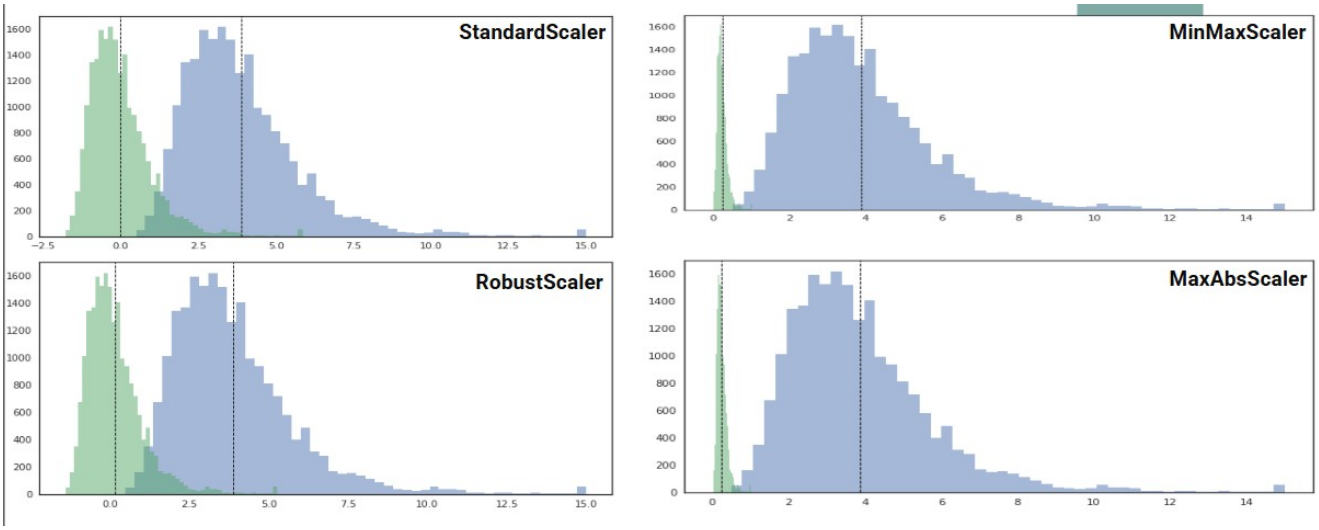
Мета: Підтвердження якості та послідовності даних

Техніки: Використання встановлених технік РАД, таких як статистичні узагальнення, візуальна перевірка та аналіз кореляції.

Техніки попередньої обробки даних

Методи трансформації даних

Назва методу	Формула	Результат	Умови використання
StandardScaler	$X^{std} = \frac{(X - \bar{X})}{\sigma_x}$	Середнє значення 0, стандартне відхилення 1	Нормальний закон розподілу
MinMaxScaler	$X_{\text{minmax}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	Масштабовані значення знаходяться в межах бажаного діапазону, зазвичай [0, 1]	Необхідність забезпечити певні межі діапазону змінних та зберегти форму розподілу (використовується майже завжди)
RobustScaler	$X_{\text{robust}} = \frac{X - \text{median}(X)}{Q_3 - Q_1}$	Масштабування засновано на значеннях міжквартільного діапазону (IQR). Викиди мають менший вплив	Використовується коли є певні викиди в даних
MaxAbsScaler	$X_{\text{maxabs}} = \frac{X}{X_{\text{absmax}}}$	Масштабовані значення знаходяться в межах діапазону [-1, 1]	Використовується коли дані мають середнє значення близько 0, розріджені дані



Візуалізація методів трансформації

Методи кодування якісних ознак:

Метод	**Умови використання**	**Результат**
One-Hot	1. Категоріальні дані, які не мають природної впорядкованості 2. Кількість категорій відносно мала	Вектор бінарних змінних
Label	1. Категоріальні дані, які мають природну впорядкованість 2. В подальшому використовується методи на основі дерев	Цілі числа замість категорій
Target	Велика кількість унікальних значень категорій	Вектор ймовірностей (частот) спостереження кожного класу для кожного унікального значення категорії, або середнє значення залежної змінної

One-Hot

	Area	Year	Item	Cassava	Maize	Plantains and others	Potatoes	Rice, paddy	Sorghum	Soybeans	Sweet potatoes	Wheat	Yams
0	Albania	1990	Maize	0	1	0	0	0	0	0	0	0	0
1	Albania	1990	Potatoes	0	0	0	1	0	0	0	0	0	0
2	Albania	1990	Rice, paddy	0	0	0	0	1	0	0	0	0	0
3	Albania	1990	Sorghum	0	0	0	0	0	1	0	0	0	0
4	Albania	1990	Soybeans	0	0	0	0	0	0	1	0	0	0

Target encoding

	Area	Year	Item	Items_target_enc
0	Albania	1990	Maize	36310.07
1	Albania	1990	Potatoes	199801.55
2	Albania	1990	Rice, paddy	40730.43
3	Albania	1990	Sorghum	18635.78
4	Albania	1990	Soybeans	16731.09

Label Encoding

	Area	Year	Item	Items_LE
0	Albania	1990	Maize	3
1	Albania	1990	Potatoes	9
2	Albania	1990	Rice, paddy	4
3	Albania	1990	Sorghum	1
4	Albania	1990	Soybeans	0

Порівняння методів кодування ознак

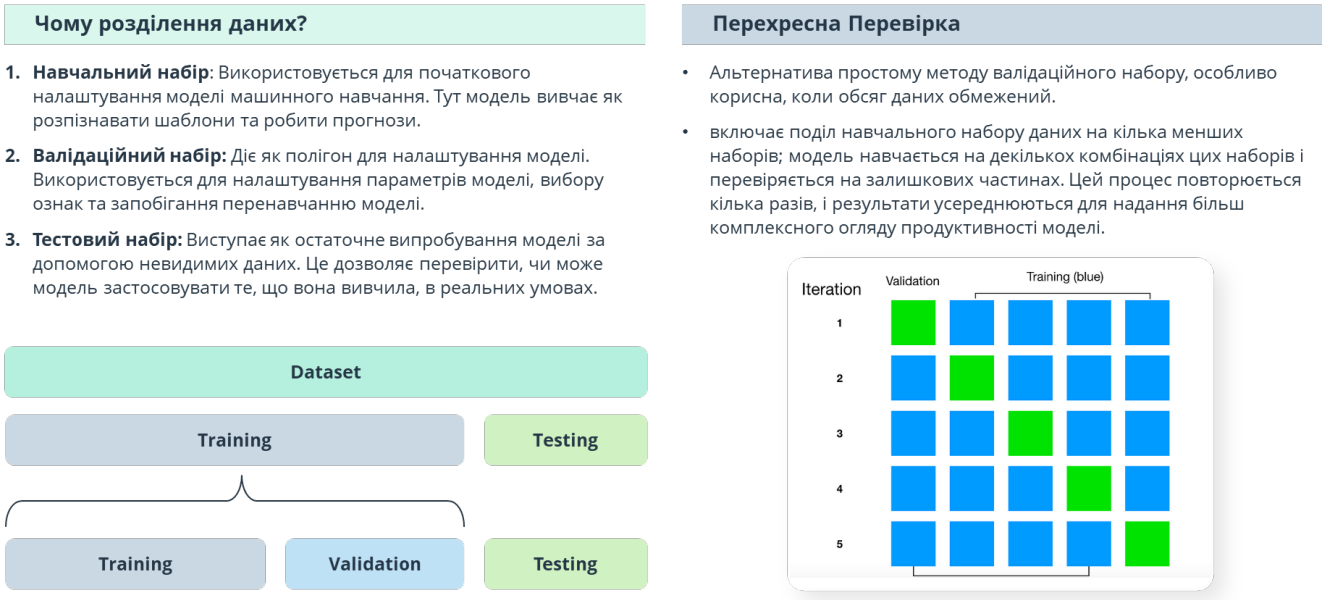
Методи вибору признаків для побудови моделі

Метод	**Способи втілення**
Фільтрація	прибирання ознак із майже унікальними або майже постійними значеннями
Статистичний відбір	Для задач регресії: • Коефіцієнт кореляції Пірсона (`f_regression`) • Mutual information (`mutual_info_regression`) $\mathrm{MI}(X,Y) = \iint p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\,dx\,dy$ Для задач класифікації: • Chi-2 тест (`chi2`) • ANOVA F-значення (`f_classif`) • Mutual information (`mutual_info_classif`) $I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$
Вибір на основі моделі	L1 регуляризація, Feature Importance, Feature Permutation

Забезпечення надійності моделі за допомогою розділення даних

Забезпечення надійності моделі є критично важливим етапом у процесі машинного навчання, і один із ключових аспектів цього процесу — розподіл даних на три основні частини: навчальну (train), валідаційну (validation) та тестову (test) вибірки.

Навчальна вибірка використовується для тренування моделі, на основі якої вона навчається розпізнавати закономірності в даних. Валідаційна вибірка служить для налаштування гіперпараметрів моделі і визначення її продуктивності під час навчання, що дозволяє відстежувати можливість перенавчання. Тестова вибірка використовується для остаточної оцінки моделі після її навчання і налаштування, надаючи дієву оцінку її здатності до генералізації на нових, невідомих даних. Додатково, використовування крос-валідації дозволяє ефективніше оцінити модель, розбиваючи дані на кілька підвибірок і проводячи навчання/тестування на різних їх комбінаціях. Це забезпечує більш стабільні оцінки моделей, оскільки дозволяє максимально використовувати доступні дані і зменшує ризик випадкових похибок, пов'язаних з наявними вибірками. Завдяки продуманому розподілу даних, можна досягти оптимізації продуктивності моделі та підвищити її надійність в умовах реального використання.



Забезпечення надійності моделі за допомогою розділення даних

Коли класи в навчальному наборі даних представлені нерівномірно у задачах класифікації виникає проблема незбалансованості спостережень. Це означає, що одна або декілька категорій мають значно більше прикладів, ніж інші. Наприклад, у

задачі виявлення хвороб рослин, приклади здорових культур можуть суттєво перевершувати хворі, що призводить до важливих проблем. Моделі, навчальні на таких даних, можуть показувати високу загальну точність, однак не матимуть адекватної здатності розпізнавати менш представлені класи, що може призвести до серйозних наслідків у застосуванні.

Для вирішення цих проблем використовують спеціальні способи підготовки та навчання моделей на незбалансованих даних:

1. Перебалансування вхідних даних (Oversampling) - цей метод полягає в збільшенні кількості спостережень менш представленого класу шляхом випадкового копіювання прикладів або генерації нових (наприклад, з використанням SMOTE — Synthetic Minority Over-sampling Technique). Це дозволяє моделі отримати більше інформації про труднощі розпізнавання меншого класу.

2. Зменшення вхідних даних (Undersampling) — цей підхід полягає у зменшенні кількості спостережень більш представленого класу, щоб збалансувати навчальний набір. Хоча цей метод забезпечує збалансованість даних, він не завжди є оптимальним, оскільки може призвести до втрати важливої інформації.

3. Зважування помилки в процесі навчання - під час навчання моделі можна використовувати зважені функції втрат, при цьому присвоюючи більшу вагу помилкам для менш представлених класів. Це дозволяє моделі зосереджуватися на покращенні точності в розпізнаванні рідкісних класів без зменшення загальної продуктивності.

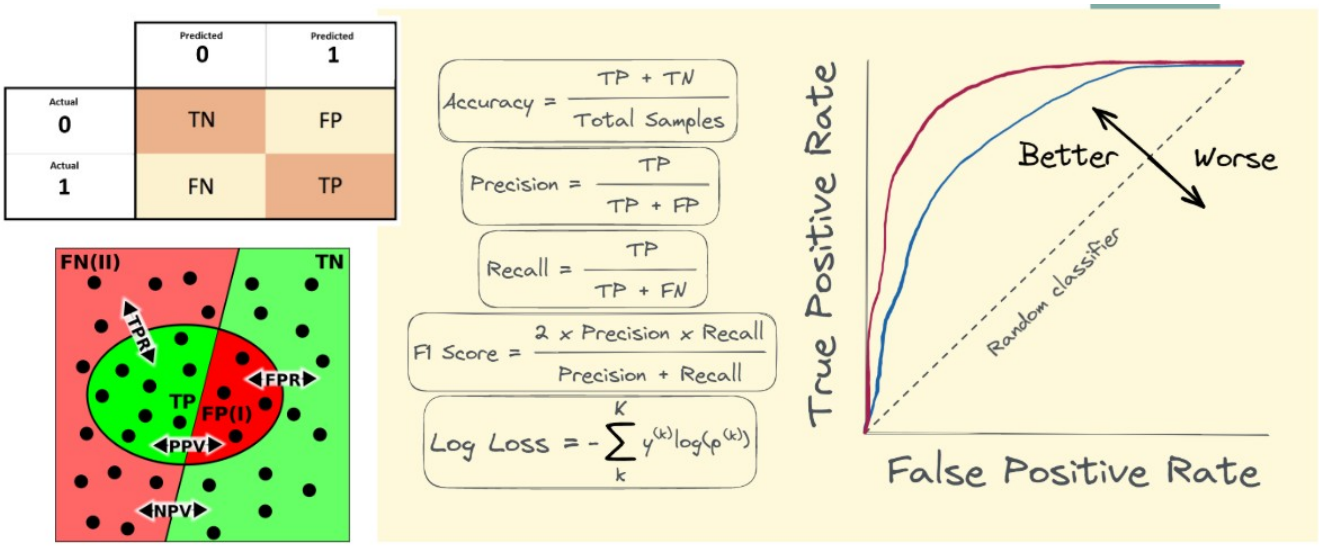
4. Вибір спеціальних метрик оцінки якості моделі - для оцінки моделей, навчених на незбалансованих даних, важливо використовувати спеціальні метрики, які відображають справжню продуктивність моделі. Приклади таких метрик включають:

Precision (точність) - відсоток правильних позитивних прогнозів з усіх позитивних прогнозів (менш представлених класів).

Recall (повнота) - відсоток правильно виявлених позитивних випадків з усіх справжніх позитивних (менш представлених класів).

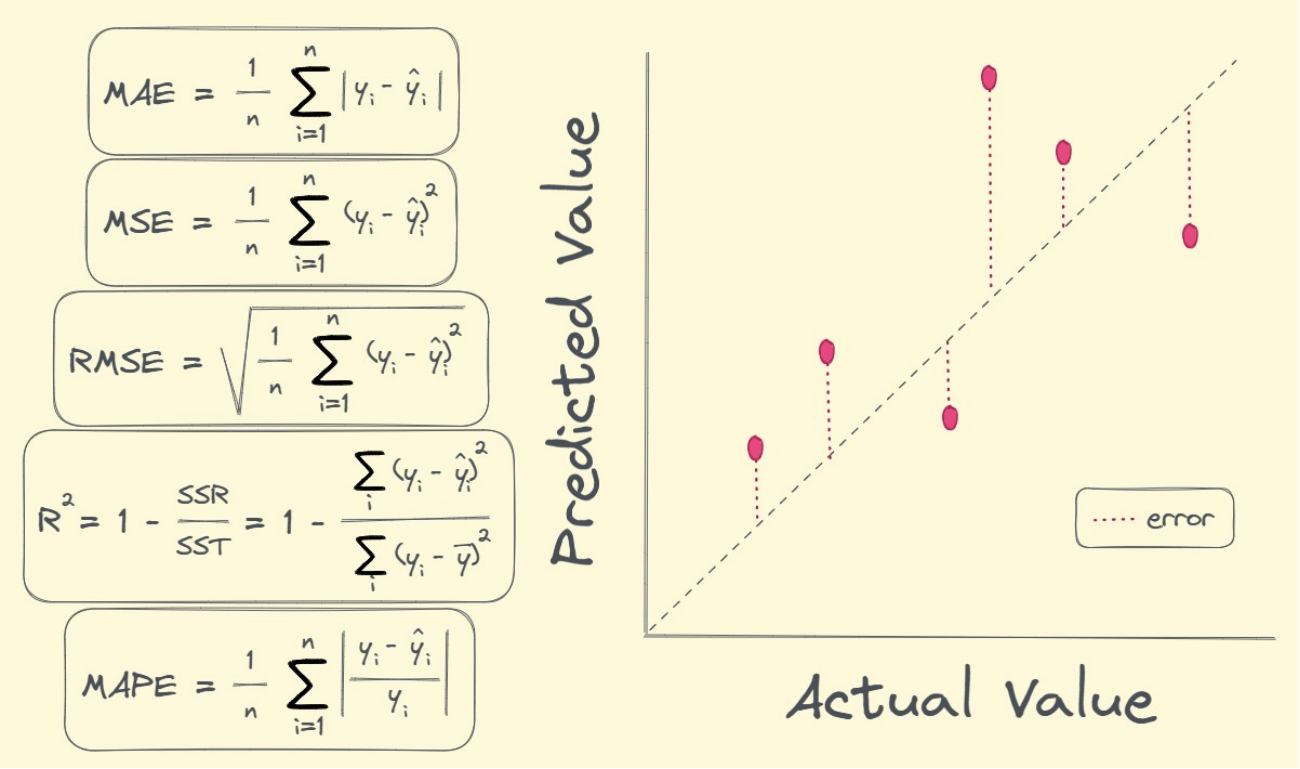
F1-міра - гармонічне середнє між точністю і повнотою, що забезпечує баланс між цими двома метриками.

ROC-AUC - Графічне представлення співвідношення між справжньо позитивними та хибно позитивними результатами, що дозволяє оцінити якість класифікації.



Метрики оцінки якості моделей (класифікація)

Метрики оцінки якості моделей регресії переважно засновані на оцінці відхилень між спостережуваними та прогнозованими значеннями. Основні метрики включають:



Середня абсолютна помилка (MAE - Mean Absolute Error) - це середнє значення абсолютних різниць між фактичними (справжніми) та прогнозованими значеннями. MAE вимірює, наскільки далеко, у середньому, прогнози відправляються від справжніх значень. Використовуйте MAE, коли важливо мати однорідну, інтуїтивно зрозумілу метрику, оскільки вона вимірює помилки в тих же одиницях, що й вихідні дані.

Середня квадратична помилка (MSE - Mean Squared Error) - це середнє значення квадратів різниць між фактичними та прогнозованими значеннями. MSE є чутливим до великих помилок, оскільки використовуються квадрати помилок. Використовуйте MSE, коли важливо більше акцентувати увагу на більших помилках. Це може бути корисним у випадках, коли великі помилки мають серйозні наслідки.

Корінь середньої квадратичної помилки (RMSE - Root Mean Squared Error) - це квадратний корінь з MSE, що повертає метрику до тих же одиниць, що й вихідні дані, отже вона оцінює продуктивність різних моделей в одних і тих же одиницях, при цьому залишається чутливою до великих відхилень, як і MSE.

Коефіцієнт детермінації (R^2) - ця метрика показує частку варіації залежної змінної, яка може бути пояснена незалежними змінними у моделі. Значення R^2 варіюється від 0 до 1, де 1 означає, що модель повністю пояснює дані. Використовуйте R^2 для оцінки загальної якості моделі. Однак, він може бути оманливим, якщо не врахувати кількість параметрів у моделі.