РОЗДІЛ 3. ПРАКТИКО-ОРІЄНТОВАНІ ТЕХНОЛОГІЇ ТА ПРИКЛАДНІ АСПЕКТИ ПІДГОТОВКИ ФАХІВЦІВ ІЗ СИСТЕМНОГО АНАЛІЗУ ТА НАУКИ ПРО ДАНІ

Лівінська Ганна Володимирівна, к.ф.-м.н, доцент
Розора Ірина Василівна, д. ф.-м.н, професор
Пономарьов Вадим Дмитрович, к.ф.-м.н, доцент
Моренець Володимир Ігорович, к.ф.-м.н
Сидоров Микола Володимир-Станіславович, к.ф.-м.н, доцент

МІЖДИСЦИПЛІНАРНА ІНТЕГРАЦІЯ МЕТОДІВ ВІЗУАЛІЗАЦІЇ ДАНИХ ТА АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВЕРИФІКАЦІЇ ЗАДАЧ КЛАСИФІКАЦІЇ

Робота присвячена методології об'єднання візуально-аналітичних підходів та алгоритмічних рішень машинного навчання (Machine Learning). Основна увага приділяється тому, як інтегроване використання візуалізації на кожному етапі розробки моделі класифікації дозволяє виявити приховані закономірності, верифікувати рішення ШІ та мінімізувати ризики, пов'язані з помилковими прогнозами у критично важливих галузях (на прикладі діагностики цукрового діабету). У статті розглядається комплексний підхід до побудови моделі машинного навчання для діагностики діабету. Основна увага приділяється не лише точності алгоритмів, а й методам візуальної верифікації на кожному етапі від очищення даних до оцінки ризиків помилкових прогнозів. Такий підхід дозволяє мінімізувати ефект «чорної скриньки» та підвищити довіру до автоматизованих систем класифікації, зокрема для медичної діагностики.

Ключові слова: візуалізація даних, машинне навчання, класифікація, верифікація моделей МН

The work is devoted to the methodology of combining visual-analytical approaches and algorithmic solutions of machine learning. The main attention is paid to how the integrated use of visualization at each stage of developing a classification model allows to reveal hidden patterns, verify AI solutions and minimize the risks associated with erroneous predictions in critical areas (for example, diabetes diagnostics). The article considers a comprehensive approach to building a machine learning model for diabetes diagnostics. The main attention is

paid not only to the accuracy of algorithms, but also to visual verification methods at each stage from data cleaning to assessing the risks of erroneous predictions. This approach allows to minimize the "black box" effect and increase trust in automated classification systems, in particular for medical diagnostics.

Keywords: data visualization, machine learning, classification, verification of ML models

1. Вступ

Розвиток штучного інтелекту характеризується стрімким ускладненням архітектур алгоритмів. Однак, паралельно із ростом обчислювальної потужності, постає фундаментальна проблема, пов'язана з втратою прозорості прийняття рішень, відома як проблема «чорної скриньки». У критично важливих галузях, таких як медицина, використанні моделей машинного навчання (МН-моделі) без візуального контролю створює високі ризики, оскільки розробник або лікар бачать лише результат (прогноз), не розуміючи логіки, що стоїть за алгоритмом прийняття рішення. Для багатьох випадків, зокрема для медичних застосувань МН-моделей, критично важливо розуміти, на основі яких факторів та співвідношень модель приймає рішення та які чинники мають найбільший вплив.

Концепція інтеграції когнітивних наук, візуальної аналітики та машинного навчання сьогодні постає як критично важливий методологічний стандарт. Необхідність такого поєднання зумовлена кількома факторами:

1. Доповнення обчислювальної потужності: Алгоритми здатні обробляти величезні масиви багатовимірних даних та знаходити математичні закономірності, які неможливо виявити вручну. Проте вони позбавлені контекстного розуміння природи даних.

2. Використання когнітивного потенціалу: Людський мозок володіє унікальною еволюційною здатністю до миттєвого розпізнавання образів, виявлення аномалій та інтерпретації складних візуальних стимулів, що виходять за межі чистої статистики.

3. Верифікація логіки: Візуалізація виступає мостом, який дозволяє експерту переконатися, що модель базується на реальних показниках та зв'язках, а не на випадкових кореляціях чи артефактах збору даних.

4. Інструментарій для оптимізації та коригування: Візуальні методи є незамінними для оцінки якості самої моделі та її тонкого налаштування. Наприклад, використання кривих Precision-Recall або аналіз розподілу ймовірностей дозволяє дослідникові свідомо коригувати поріг класифікації. Це дає змогу гнучко адаптувати модель під конкретну медичну задачу, зокрема

мінімізувати пропуски захворювань (збільшення Recall) або уникнути хибних діагнозів (збільшення Precision), що неможливо зробити ефективно, спираючись лише на усереднені метрики на кшталт Accuracy.

При побудові ML-моделей для розв'язанні задач класифікації, зокрема у діагностиці критичних захворювань в медицині, ціна помилки виходить далеко за межі статистичних метрик. Помилки класифікації призводять до таких ризиків:

- Етичні ризики: Хибнопозитивний результат викликає необґрунтований стрес у пацієнта та призначення зайвого лікування, що може мати побічні ефекти.
- Безпекові ризики: Хибнонегативний результат (пропуск хвороби) може призвести до прогресування захворювання, критичних ускладнень або смерті через відсутність своєчасної терапії.
- Фінансові втрати: Неефективне використання ресурсів медичних закладів на додаткові обстеження та лікування ускладнень, яких можна було запобігти.

Дослідник має зважати на наявність можливих ризиків для конкретної задачі, визначити та оцінити вартість помилки класифікації того чи іншого роду та будувати модель з урахуванням цих деталей, а не покладаючись на стандартні установки. Саме тому візуальна верифікація виступає не просто як допоміжний інструмент, а як необхідний фільтр безпеки та інструмент точного налаштування, що мінімізує ризики помилкових прогнозів у системах підтримки прийняття рішень.

Таким чином, інтеграція методів візуалізації на кожному етапі розробки дозволяє перетворити «чорну скриньку» на прозору систему. Це забезпечує поєднання математичної точності обчислень із фаховою інтуїцією та досвідом людини.

Проблема інтерпретованості моделей (eXplainable AI, XAI) активно досліджується в останнє десятиліття. У фундаментальних роботах [1], [2] було закладено математичні основи локальної інтерпретації прогнозів. В багатьох роботах підкреслено, що самих лише математичних значень недостатньо для прийняття рішення, необхідне інтерактивне середовище, яке дозволяє користувачу «спілкуватися» з даними (див., наприклад, [3], [4]).

У медичній сфері питання верифікації моделей розглядалися в численних працях, присвячених аналізу діабетичних даних, онкологічних даних, тощо. Автори підкреслюють, що наявність пропущених значень та біологічних викидів у таких наборах даних часто призводить до викривлення

результатів, якщо не використовувати методи візуального розвідувального аналізу (див., напр., [4]).

Метою цієї роботи є демонстрація важливості комплексного підходу до розробки та верифікації моделі класифікації на прикладі задачі діагностики цукрового діабету на базі відкритого датасету Pima Indians Diabetes Dataset. Показано, як візуальні методи допомагають не лише зрозуміти модель, а й покращити її якість на кожному етапі. Зауважимо, що в даній роботі ми розглядаємо датасет, який містить значення ознак, виражені кількісними змінними. Для роботи з якісними змінними використовуються дещо інші діагностичні діаграми, проте візуальний супровід при побудові моделі також є надзвичайно важливим.

2. Візуалізація як фундамент підготовки даних

Перш ніж приступати до навчання моделі класифікації, необхідно провести ретельний аудит даних. У цьому розділі ми розглянемо, як візуальний аналіз допомагає виявити приховані вади у структурі даних, що можуть критично вплинути на надійність прогнозу. На відміну від сухої статистики, візуалізація на цьому етапі дає фундаментальні ідеї для подальшої обробки та запобігає навчанню моделі на некоректних даних.

2.1. Датасет 'PimaIndiansDiabetes2'

Для демонстрації методології обрано класичний набір даних Національного інституту діабету, хвороб травлення та нирок (NIDDK). Вибірка включає 768 пацієнток (жінок індіанського походження племені Піма віком від 21 року) і містить 8 діагностичних предикторів (таких як рівень глюкози, ІМТ, артеріальний тиск, кількість вагітностей, тощо) та одну цільову змінну -- наявність діабету (Рис. 1). Наша задача -- передбачити, чи є у жінки з вказаним набором ознак (1-8) діабет (значення цільової змінної pos) чи немає (neg).

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
1	6	148	72	35	NA	33.6	0.627	50	pos
2	1	85	66	29	NA	26.6	0.351	31	neg
3	8	183	64	NA	NA	23.3	0.672	32	pos
4	1	89	66	23	94	28.1	0.167	21	neg
5	0	137	40	35	168	43.1	2.288	33	pos
6	5	116	74	NA	NA	25.6	0.201	30	neg
7	3	78	50	32	88	31.0	0.248	26	pos
8	10	115	NA	NA	NA	35.3	0.134	29	neg
9	2	197	70	45	543	30.5	0.158	53	pos
10	8	125	96	NA	NA	NA	0.232	54	pos
11	4	110	92	NA	NA	37.6	0.191	30	neg
12	10	168	74	NA	NA	38.0	0.537	34	pos
13	10	139	80	NA	NA	27.1	1.441	57	neg
14	1	189	60	23	846	30.1	0.398	59	pos
15	5	166	72	19	175	25.8	0.587	51	pos

Рис.1. Змінні датасету 'PimaIndiansDiabetes2'

Зауважимо, що стартовий етап роботи з даними містить зокрема обробку пропущених значень. Оскільки метою даної статті не є методики роботи з пропущеними даними, отже ми просто видалили неповні рядки. В загальному випадку, звичайно, є ряд методик імпутації, які дозволяють мінімізувати втрати корисної інформації на цьому етапі. Такий підхід гарантує, що модель буде навчатися на біологічно коректних закономірностях, а не на артефактах обробки даних.

2.2. Аналіз топології даних: виявлення класів та зон перекриття

Для візуалізації структури багатовимірних даних та в нашому випадку визначення того, як клінічні показники формують простір станів пацієнтів, варто провести аналіз головних компонент (PCA, див. Рис. 2,3). Цей підхід дозволяє знизити розмірність даних, зберігши при цьому більшу частину загальної дисперсії в межах перших двох компонент.

В нашому випадку перші дві головні компоненти (Dim1 та Dim2) сумарно пояснюють 51.45% загальної варіативності даних (31.99% та 19.46% відповідно). Хоча це значення свідчить про високу складність даних, воно є достатнім для візуалізації основних закономірностей розподілу пацієнтів на площині. Проте в загальному випадку для подальшого покращення моделі варто розглянути аналіз залишку дисперсії (48.5%), який не потрапив на графік, можливо, шляхом додавання третьої компоненти або використання інших методів зниження розмірності.

Графічна візуалізація змінних в площині головних компонент (Рис. 2, Correlation Circle) потрібна для демонстрації того, які саме (біологічні) чинники стоять за геометричним розподілом точок датасету на площині головних компонент (Рис. 3). Діаграма демонструє ортогональність певних клінічних векторів, що свідчить про їх незалежність у межах даної патології. Довжина стрілки вказує на якість змінної на площині. Чим ближче стрілка до кола, тим краще вона пояснюється цими двома компонентами. Якщо стрілки знаходяться близько одна до одної, ці змінні сильно корелюють. Якщо стрілки направлені під кутом 90 градусів, відповідні їм змінні не пов'язані між собою.

Для нашого датасету на Рис. 2 бачимо, що змінні glucose (рівень глюкози), mass (IMT), triceps (товщина шкірної складки), age (вік) мають сильний позитивний зв'язок із першою головною компонентою (Dim1). Це вказує на те, що Dim1 відображає загальний метаболічний профіль. Змінні age (вік) та pregnant (кількість вагітностей) сильно корелюють між собою (відповідні їм стрілки знаходяться близько одна до одної) та мають значний внесок у від’ємні значення другої головної компоненти (Dim2). Саме вони спричинюють зміщення групи pos (діабет) в праву частину графіка, зображеного на Рис. 3. Glucose має найдовшу проєкцію на першу компоненту. Це означає, що вона робить найбільший внесок у розділення людей на групи за наявністю діабету. Показник age розташований майже перпендикулярно до вектора mass, що вказує на слабку кореляцію між цими параметрами в даній вибірці.

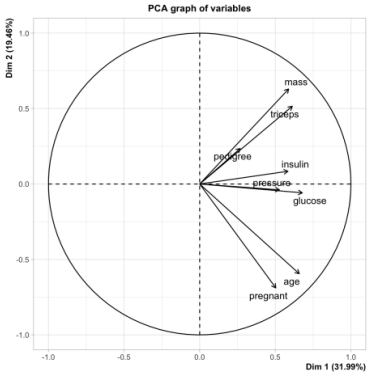


Рис. 2. Діаграма навантажень головних компонент датасету 'PimaIndiansDiabetes2'

Біплот (Рис. 3) – це графік, який відображає проєкцію багатовимірних даних на площину головних компонент. Аналіз біплоту, де точки марковані відповідно до фактичного діагнозу, дозволяє ідентифікувати наступні топологічні особливості:

1. Класова сегрегація: Пацієнти без діабету (клас 0 або neg) переважно концентруються в області нижчих значень першої головної компоненти (Dim1). Вектор глюкози та індексу маси тіла спрямований у протилежний бік, що підтверджує їх роль як домінантних факторів ризику. Попри перекриття, ми бачимо чітке зміщення: група neg тяжіє до лівої частини графіка (негативні значення Dim1); Група pos (діабет) зміщена вправо та трохи вище/нижче по осі Dim2. Еліпс neg більш компактний і витягнутий вздовж горизонтальної осі. Це свідчить про те, що показники здорових людей більш однорідні. Еліпс pos значно ширший. Це означає, що група діабету має велику варіативність показників (високий рівень глюкози, тиску або ІМТ може проявлятися по-різному в різних комбінаціях).

2. Зони високої ентропії (перекриття): Спостерігається значна область перекриття класів у центрі координатної площини. Це свідчить про існування "перехідних станів" – пацієнтів, чий біохімічний профіль уже наближається до патологічного, але ще не досяг клінічного порогу встановлення діагнозу, і, зокрема, про те, що діагноз неможливо встановити лінійно за однією або двома ознаками, що обґрунтовує необхідність використання більш складних методів машинного навчання (наприклад, випадкового лісу або нейронних мереж). Перекриття кластерів є очікуваним для медичних даних, де показники здоров'я (глюкоза, тиск, ІМТ) часто мають неперервний характер і не завжди утворюють ізольовані групи.

3. Атипові спостереження: Виявлено окремі випадки класу 1 (діабет) у зоні норми, що може вказувати на специфічні форми захворювання, не пов'язані з класичними факторами ожиріння чи віку. На графіку видно кілька жовтих трикутників (pos) у верхньому правому куті. Це пацієнти з критично високими показниками, які вибиваються із загальної маси навіть серед хворих.

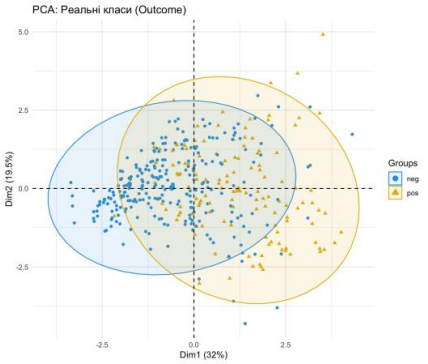


Рис. 3. Біплот для датасету 'PimaIndiansDiabetes2'

Отже, дані мають структуру, що дозволяє відрізнити хворих від здорових (зміщення груп вправо вздовж Dim1), але через велику зону перекриття у центрі графіка, діагностика лише за цими ознаками потребує більш складних методів аналізу, ніж простий візуальний поділ. Наявність розмитих меж та зон перекриття між класами підкреслює нелінійну природу досліджуваного явища. Топологічна складність даних обґрунтовує використання ансамблевих методів машинного навчання, здатних вибудовувати складні гіперплощини розділення там, де класичні лінійні методи дають похибку через перекриття ознак.

2.3. Інтеграція статистичних методів та графічного аналізу. Виявлення прихованих аномалій

Для глибшого розуміння природи даних та валідації якості вибірки необхідно провести розвідувальний аналіз даних (EDA) з використанням методів дескриптивної статистики та візуалізації. Самі по собі статистичні показники (середнє, медіана, дисперсія) можуть бути оманливими без графічного підтвердження.

Ключовими зображеннями на цьому етапі є діаграми розсіювання даних по класах, боксплоти, гістограми, розподіл даних по класах. Їх можна зобразити окремо для більшої деталізації зображення, інколи зручно подивитись на всі діаграми, виведені в одному графічному вікні, одним з таких комбінованих графіків є варіація матричної діаграми розсіювання (Matrix Plot), таке зображення є надзвичайно інформативним, оскільки

поєднує в собі одразу чотири типи аналізу: розподіли, кореляції, попарні залежності та порівняння груп (Рис. 4).

2.3.1. Аналіз розподілів

Використання гістограм та діаграм розсіювання дозволяє оцінити характер розподілу кожної змінної, що важливо зокрема для побудови моделей, які вимагають нормального розподілу даних. Окрім того, при зображенні на одному графіку гістограм для різних класів, можна побачити різницю в розподілі ознак для різних класів, що є ознакою того, що такі змінні будуть ефективними при включенні в модель класифікації.

На діаграмі розподілу цільової змінної часто спостерігається незбалансованість класів, коли кількість спостережень одного з класів суттєво (інколи критично) перевищує кількість спостережень іншого класу. Така структура даних створює загрозу "перекосу" моделі в бік мажоритарного класу. Алгоритм може демонструвати високу загальну точність (Accuracy), проте мати низьку чутливість (Sensitivity) до виявлення реальних випадків, наприклад, рідкісного захворювання. Ця проблема обумовлює необхідність використання спеціальних метрик оцінки (F1-score, AUC-ROC), корекцію значення порогової ймовірності класифікації та, можливо, методів ресемплінгу (SMOTE) на етапі підготовки даних.

2.3.2. Порівняльний аналіз груп та візуалізація викидів

Сирі дані (зокрема медичні) часто містять екстремальні (аномальні) значення, які, якщо їх лишити в датасеті, часто дуже псують результати та точність аналізу. Статистичні середні показники приховують цю проблему, проте графічний аналіз дозволяє відрізнити фізіологічно можливі викиди (наприклад, дуже високий ІМТ) від технічних помилок запису. Зокрема, часто може бути виявлено нульові значення у критичних параметрах (ІМТ, тиск), що з технічної (медичної) точки зору свідчить про пропуски в даних (missing values), замасковані під нулі, а не на реальні показники, і потребує подальшої імпутації (заповнення пропусків).

Боксплот (Box-and-whisker plot) – це простий, але дуже інформативний графік, який дає можливість побачити, якщо зображений для однієї вибірки, симетричність розподілу ознаки, розподіл елементів вибірки по квартилях, наявність викидів (аномальних спостережень). Окрім того, при зображенні боксплотів для різних груп поруч може бути виявлена міжгрупова диференціація, якщо помітна суттєва різниця між характеристиками центру та розсіювання ознак в різних групах. Для формального виявлення

статистичної значущості різниці між групами треба проводити додатково статистичні тести та оцінку розміру ефекту, проте візуальна оцінка дозволяє швидко виявити можливі ознаки, значення яких для різних класів є значущо різними. Це свідчить про високу роздільність класів за відповідною ознакою.

2.3.3. Аналіз кореляцій та статистичних залежностей

Теплові карти (Heatmap) та інші зображення кореляційних матриць є ключовим інструментом для розуміння структури залежності даних. Вони допомагають виявити таку проблему як мультиколінеарність ознак. Наприклад, сильна кореляція між віком та кількістю вагітностей може призвести до перекосу ваг у лінійних моделях та загальної нестійкості (ефект перенавчання) моделей.

За виявлення такої проблеми, особливо у випадку, коли планується побудова регресійних моделей, варто формально перевірки наявність мультиколінеарності, наприклад, через обчислення коефіцієнтів здуття дисперсії (VIF) для кожної незалежної змінної.

З іншого боку, ознаки, які мають незначущий зв'язок з цільовою змінною, можуть бути виключені з моделі. Часто це призводить до спрощення моделі без втрати точності (інколи навіть точність підвищується за рахунок відсутності зайвих змінних, які заважають приймати правильне рішення), підвищення інтерпретованості моделі, підвищенню швидкості роботи алгоритму.

Встановлені статистичні закономірності стають фундаментом для вибору методів попередньої обробки даних (масштабування та обробка нулів) перед етапом моделювання.

2.3.4. Попарні діаграми розсіювання

На попарних діаграмах розсіювання (нижній трикутник діаграми на Рис. 4), де об'єкти, що належать різним класам, зображуються різними кольорами, слід звернути увагу перш за все на перекриття класів. Якщо наявне помітне розділення класів по певній змінній, ця змінна буде корисною в моделі класифікації. Часто наявні змінні, для яких розподіл по класах виглядає рівномірним. Швидше за все, така змінна або не буде корисною для моделі, або навіть буде псувати модель класифікації.

Також на діаграмах розсіювання часто можна побачити інші проблеми. Наприклад, можливе скупчення точок вздовж осей свідчить про типову проблему "нульових значень", які потребують обробки перед навчанням моделі.

Хоча загальна матриця розсіювання дає комплексне уявлення про дані, висока щільність точок часто маскує локальну структуру розподілу цільової змінної. Для глибшого аналізу сепарабельності класів доцільно розглянути проєкції багатовимірного простору на площини найбільш інформативних ознак, наприклад, в нашому випадку, glucose-age та glucose-mass (Рис. 5). Така декомпозиція дозволяє чіткіше візуалізувати межу роздільності класів (decision boundary) та виявити області концентрації позитивних випадків діабету, які в двовимірному просторі формують виразні кластери. Окремий вивід цих графіків підтверджує, що комбінація високого рівня глюкози з віком або індексом маси тіла є ключовим нелінійним предиктором, що значно спрощує подальшу інтерпретацію результатів моделювання.

2.3.5. Розвідувальний аналіз для 'PimaIndiansDiabetes2'

В нашому випадку (Рис. 4) матриця візуалізації дозволяє провести глибокий аналіз структури даних, виділяючи ключові закономірності між фізіологічними показниками та наявністю діабету. Гістограми на діагоналі дають можливість провести аналіз розподілів. Більшість показників (insulin, pedigree, age, pregnant) мають виражений правосторонній скос. Це означає, що у вибірці переважають молоді люди з низькими показниками інсуліну, а екстремально високі значення зустрічаються рідше. Частина ознак мають розподіл, близький до нормального. Помітний зсув гістограм змінних для різних класів glucose, age, mass, triceps, insulin свідчать про вищі в середньому показники в групі з діабетом.

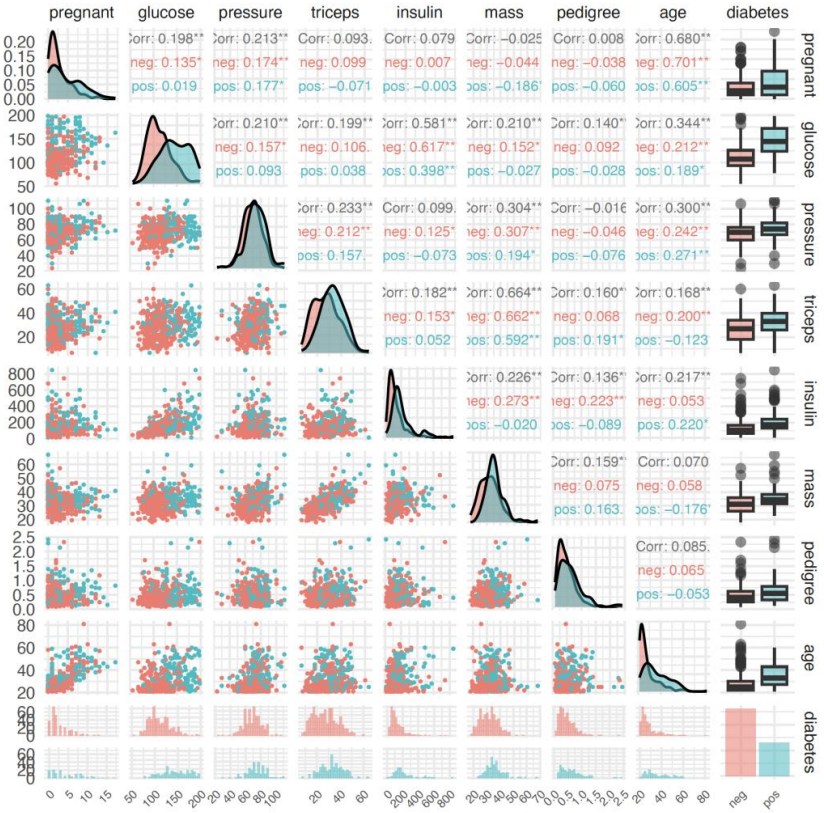


Рис. 4. Діаграма розсіювання для 'PimaIndiansDiabetes2' з розподілом на класи за всіма наявними в датасеті змінними

Розподіл цільової змінної (diabetes) показаний на стовпчиковій діаграмі в нижньому правому кутку. Вона каже про суттєву незбалансованість класів. Кількість пацієнтів без діабету (neg) майже вдвічі перевищує кількість хворих (pos). Це критичний момент для вибору метрик оцінки моделі (пріоритет на Precision/Recall, або інших критичних для конкретної даної задачі метрик, але не Accuracy). Окрім того, для покращення моделі варто в загальному випадку скористатися додатково методиками, які дозволяють збалансувати класи.

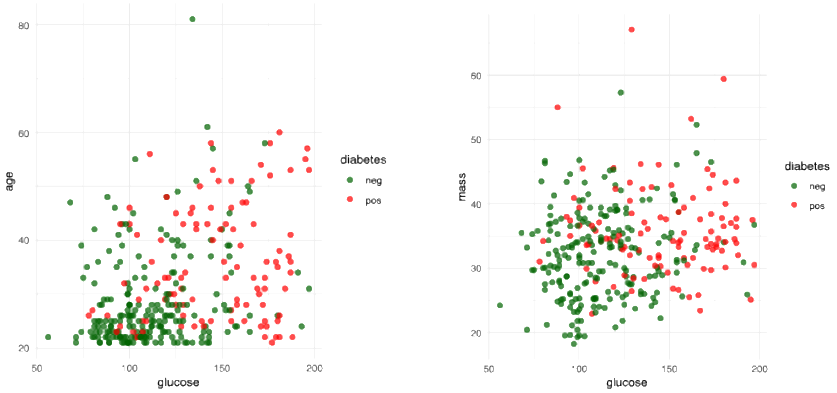


Рис. 5. Діаграма розсіювання з розподілом на класи в площині змінних “глюкоза” - “вік” (ліворуч) та “глюкоза” - “індекс маси тіла” (праворуч)

Аналіз боксплотів (правий стовпчик на Рис. 4) виявив також наявність викидів у багатьох показниках, зафіксовано значення, що виходять за межі 1.5 міжквартильного розмаху. Велика кількість точок за межами "вусів" у змінних insulin та pregnant вказує на наявність специфічних клінічних випадків або аномалій у зборі даних.

Боксплоти, побудовані в розрізі класів (здоровий/хворий), демонструють помітну різницю в медіанних значеннях, що свідчить про сепарацію класів за певними ознаками. Для змінної glucose найбільш виражена різниця між медіанами. У групи pos медіана значно вища, що підтверджує статус глюкози як головного предиктора. ІМТ та вік також показують помітний зсув медіан у бік вищих значень для групи з діабетом, що свідчить про їх значущий внесок у формування ризику.

Числові значення коефіцієнтів кореляції Пірсона (варіант за замовчуванням, верхній трикутник діаграми) виявляють важливі зв'язки в даних. Маємо свідчення мультиколінеарності ознак. Найсильніший зв'язок спостерігається між ознаками age та pregnant (0.68, природна позитивна кореляція), а також між mass та triceps (0.664, висока кореляція вказує на те, що обидва показники характеризують ступінь ожиріння). Це логічно з біологічної точки зору. Сильно значуще значення коефіцієнту кореляції між глюкозою та інсуліном (0.58) теж свідчить про очікуваний фізіологічний зв'язок, що підтверджує релевантність даних медичній теорії. Водночас слабка кореляція деяких ознак (наприклад, pressure – кров’яний тиск)

підштовхує до ідеї перевірки певних моделей, таких як дискримінантний аналіз, логістична регресія, на надмірність (overfitting).

Аналізуючи діаграми розсіювання, видно, що на багатьох графіках (наприклад, age - pressure) спостерігається сильне змішування червоних (neg) та блакитних (pos) точок. Це свідчить про те, що дані не є лінійно роздільними в базовому просторі ознак по жодній з окремо взятих змінних, проте класи все ж мають зсув по ряду змінних.

Отже, можемо зробити висновок попереднього аналізу датасету 'PimaIndiansDiabetes2'. Дані мають складну структуру з вираженим дисбалансом класів та нелінійними залежностями. Найбільшу роздільну здатність мають показники глюкози, віку та маси тіла, що підтверджується як візуальним аналізом розподілів, так і результатами кореляцій та PCA. Наявність викидів вимагає обов'язкового етапу стандартизації та імпутації даних.

Тепер маємо структурований фундамент для побудови надійного діагностичного алгоритму. Для багатьох моделей класифікаторів, таких як наївний баєсів класифікатор, випадковий ліс, всі ознаки доцільно залишити в моделі для подальшого тренування. Проте для побудови моделей на базі регресії (логістична регресія, дискримінантний аналіз) варто вилучити певні змінні для уникнення ефекту перенавчання та забезпечення стійкості моделі.

3. Побудова моделі логістичної регресії

Для прогнозування наявності діабету в якості базової моделі було обрано модель логістичної регресії. Це статистичний метод, що використовується для передбачення ймовірності події шляхом припасування даних до логістичної кривої. Модель логістичної регресії, з одного боку, дає непогані результати класифікації, а з іншого, на ній можна продемонструвати деякі додаткові діаграми, які свідчать про роздільність класів, зокрема залежно від вибору порогового рівня ймовірності класифікації. Власне, основним завданням алгоритму класифікації є знаходження оптимальної межі прийняття рішень (decision boundary), яка найкращим чином розмежовує об'єкти різних класів у просторі ознак, і для моделі логістичної регресії цю межу можна зобразити.

3.1. Математична модель

В основі моделі лежить лінійна комбінація ознак, яка доповнюється логістичним перетворенням.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_m \quad (1)$$

де β_i – коефіцієнти моделі, а x_i – значення ознак (наприклад, глюкоза, вік).

Модель повинна прогнозувати значення ймовірності приналежності об'єкта, який має вказаний набір значень ознак, до цільового класу (в нашому випадку – наявність діабету). Щоб отримати значення ймовірності в межах від 0 до 1, використовується функція сигмоїди:

$$P = \sigma(z) = \frac{1}{1+e^{-z}} \quad (2)$$

Ця функція дозволяє інтерпретувати результат P як ймовірність $P(y = 1|x)$, де y – клас об'єкта, $y \in \{0,1\}$, $x = (x_1, \dots, x_m)$ – вектор ознак.

Класифікація відбувається шляхом порівняння ймовірності з порогом (за замовчуванням це 0.5). Математично межа розділення класів в моделі логістичної регресії – це гіперплощина, для якої $\sigma(z) = 0.5$, а отже $z = 0$. Зокрема в одновимірномі випадку (коли модель будується лише на основі однієї змінної, наприклад, рівня глюкози) ми можемо зобразити безпосередньо графік сигмоїди (2) (Рис. 6), у двовимірному просторі ознак це пряма лінія, яку можна побачити на графіку (Рис. 8):

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

Проте типовою є ситуація, коли стандартний поріг ймовірності 0.5 не дає задовільних результатів класифікації, і зміна цього значення дозволяє суттєво покращити якість моделі. Візуалізація роздільності класів дозволяє попередньо оцінити якість класифікації «за замовчуванням» та можливості покращення моделі шляхом зміни порогового рівня ймовірності.

Для побудови роздільчих меж, оскільки ми не можемо зобразити модель в просторі високої розмірності, варто спочатку виділити найбільш ефективні ознаки, які для моделі логістичної регресії мають найвищу ефективність. Корисним зображенням для цього є діаграма важливості ознак.

3.2. Діаграма важливості ознак

Діаграма важливості ознак будує та демонструє коефіцієнти логістичної регресії β_i^* при кожній з нормалізованих змінних моделі, які є зручними для математичних розрахунків та статистичних тестів, і показують напрямок зміни результуючої змінної залежно від зміни значення ознаки, або ж відношення шансів (OR, Odds Ratio), більш традиційне при аналізі ризиків

в біології, медицині, тощо. Відношення шансів має просте трактування. В нашому випадку значення OR=3.25 для змінної glucose означає, що шанси на виникнення діабету зростають у 3.26 рази при зростанні рівня глюкози на одне стандартне відхилення.

Ми опускаємо математичні деталі, проте зауважимо, що при класифікації за допомогою логістичної регресії саме аналіз коефіцієнтів при нормалізованих змінних дозволяє зрозуміти, чому модель приймає те чи інше рішення.

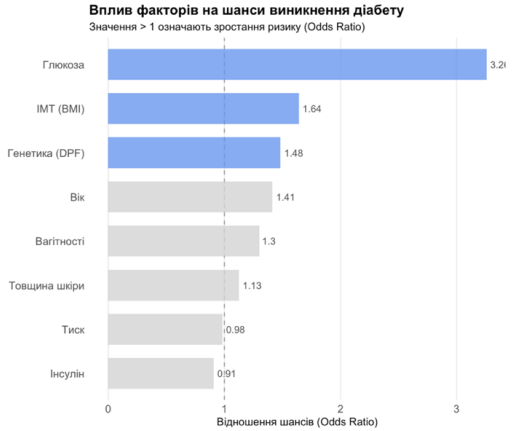


Рис. 6. Діаграма важливості ознак для класифікації наявності діабету. Відношення шансів для кожної з ознак (стандартизованих)

На Рис. 6 зображена діаграма важливості ознак для класифікації наявності діабету на базі даних датасету 'PimaIndiansDiabetes2', де демонструється значення відношення шансів для кожної з ознак (стандартизованих). Синім позначено значущі на рівні 0.05 ознаки. Згідно цього графіку глюкоза, маса та спадковість є визначальними факторами. Насправді, на рівні 0.1 значущою змінною є також вік. Всю детальну інформацію можна отримати при побудові математичної моделі та аналізу відповідного статистичного звіту.

Діаграма важливості ознак дозволяє зробити наступні висновки:

1. Глюкоза є найвпливовішою ознакою. Якщо у пацієнта високий рівень цукру, модель майже автоматично підвищує ймовірність діабету. Odds Ratio близько 3.26 означає дуже суттєвий вплив. Це пояснює, чому на

попередніх графіках ми бачили найбільш чіткий поділ саме за віссю глюкози. Навіть невеликі (стандартизовані) коливання цього показника мають вдвічі більший вплив на ризик, ніж зайва вага або генетична схильність.

2. Індекс маси тіла – другий за важливістю фактор. Він має "вагу" 1.64. Це підтверджує медичну теорію про те, що ожиріння є ключовим тригером для діабету 2-го типу.

3. Генетична схильність також є важливим фактором, коефіцієнт 1.48.

4. Вік та Кількість вагітностей. Ці фактори мають помірний вплив (OR = 1.41, 1.3). Вони додають фоновий ризик. Ми бачили, що кількість вагітностей у цьому наборі даних корелює з віком, тому їх ефекти частково перекриваються.

5. Слабкі фактори (Тиск, Інсулін): Зверніть увагу, що pressure (кров'яний тиск) знаходиться біля одиниці (0.98). Це означає, що в межах цього дослідження тиск сам по собі не є хорошим індикатором діабету. Це пояснює, чому модель може помилятися на пацієнтах з високим тиском, але нормальною глюкозою.

Визначивши найбільш впливові змінні, можемо візуалізувати межі роздільності класів по них в різних проєкціях.

3.3. Візуальний аналіз роздільності класів

Оскільки в багатовимірному просторі важко візуалізувати роботу моделі, ми звертаємося до аналізу функції сигмоїди для окремих змінних, які володіють найбільшою розділюючою здатністю, та аналізу 2D-проєкцій для двох змінних. На таких діаграмах чіткіше простежується логіка роботи моделі. Зокрема ми можемо побачити наскільки ефективною є лінійна роздільність моделі, оскільки логістична регресія проводить пряму лінію (у 2D), з метою мінімізувати помилки класифікації. Візуалізація дозволяє побачити також перехідні зони, де ймовірність віднесення до позитивного класу змінюється від низької до високої.

Використання саме цих проєкцій підтверджує, що модель спирається на ключові біологічні маркери, а графічне відображення межі розділення дозволяє оцінити якість класифікації не лише через метрики (Accuracy, Precision), а й візуально, аналізуючи кількість "проникнень" одного класу в зону іншого.

3.3.1. Візуалізація роздільності за однією ознакою

Найвпливовішою на наявність діабету, судячи з попередньої візуалізації датасету, є змінна glucose. Для демонстрації процесу класифікації

за допомогою сигмоїдної кривої зобразимо її графік. Для цього ми використаємо модель (2) лише з однією ознакою – glucose. Модель оцінює ймовірність належності до цільового класу 1 (наявність діабету) через функцію сигмоїди. Точка перетину кривої з лінією y=0.5 є стандартним критичним порогом класифікації для однієї змінної.

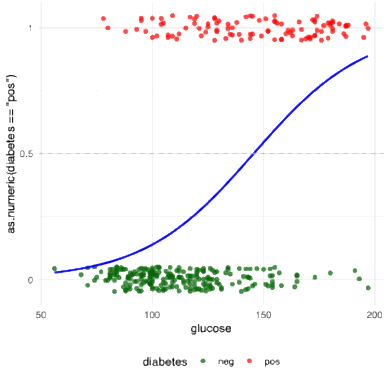


Рис. 7. Імовірність наявності діабету залежно від рівня глюкози (синя крива)

Отже, на Рис. 7 бачимо зображення моделі простої логістичної регресії, яка на основі лише однієї змінної, рівня глюкози, обчислює оцінку ймовірності наявності діабету. Як бачимо на рисунку, при низьких значеннях глюкози ймовірність переважно ближча до 0 (зелені точки neg), а при високих – до 1 (червоні точки pos). Варто звернути увагу, що у випадку встановлення стандартного порогу ймовірності для прийняття рішення щодо розділу на класи 0.5 (пунктирна пряма) ми б мали для тих жінок, які мають рівень глюкози вищий за приблизно 145, визначити наявність діабету, а тих, у кого цей показник менший, визнати здоровими. При цьому бачимо, що більшість жінок без діабету будуть правильно класифіковані як здорові, проте досить багато жінок з діабетом помилково потраплять в категорію neg, що, власне, недобре, інколи навіть критично. Тобто проста модель класифікує наявність діабету погано. Теоретично, якщо встановити пороговий рівень глюкози приблизно 80, то майже всі жінки з діабетом потраплять в позитивний клас. При цьому поріг ймовірності суттєво знизиться і більшість здорових жінок помилково потраплять в позитивний клас, що робить модель класифікації не дуже корисною.

Отже. незважаючи на те, що змінна glucose є дуже значущою при виявленні діабету, одна вона не зможе забезпечити ефективне розділення класів. Додаючи інші ознаки, ми можемо також побачити розділювальні межі у проєкції двох ознак для різних значеннях порогового рівня ймовірності.

3.3.2. Візуалізація межі розділення для двох ознак

Якщо на загальній діаграмі розсіювання структура класів виглядає занадто зашумленою, варто подивитися окремі діаграми для найбільш значущих пар ознак, як-от glucose-age та glucose-mass (Рис. 5).

Для моделі з двома предикторами ми вже не побудуємо зрозумілий графік модельної кривої, подібно до того, як ми зобразили графік логістичної кривої для моделі з одним предиктором glucose (Рис. 7), проте ми можемо побудувати інший графік, який демонструє те, як модель розділяє класи залежно від встановленого порогового рівня ймовірності приналежності до цільового класу – так звані розподільчі межі (decision boundaries).

Коли ми додаємо другу зміну (наприклад, mass), межа розділення стає лінійною функцією у просторі ознак. Математично це спирається на поняття логарифму шансів (Log-Odds), або логіту (logit).

Неважко побачити, що якщо взяти логарифм в рівнянні (2), де лінійна комбінація (1) будується лише по двох змінних, ми легко можемо отримати рівняння

$$\ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \beta_0 + \beta_1 \cdot \mathrm{glucose} + \beta_2 \cdot \mathrm{mass} \qquad (3)$$

Тут P – це ймовірність приналежності до цільового класу, яка моделюється рівнянням (2), $P/(1-P)$ – шанси (Odds), $\ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right)$ – логарифм шансів або логіт.

Отже, по суті, хоча ймовірність P змінюється нелінійно, межа прийняття рішення залишається прямою лінією, яку можна зобразити, якщо зафіксувати значення ймовірності P.

Аналіз 2D-проєкцій (наприклад, glucose проти mass) дозволяє побачити цю пряму. На графіку ми також спостерігаємо зони впевненості, де точки одного класу знаходяться далеко від межі, та область невизначеності – точки біля самої лінії, де ймовірність наближається до рівня порогу класифікації (наприклад, 0.5), що вказує на ризик помилкової класифікації через перекриття ознак.

Використання такого підходу дозволяє зрозуміти фізичний зміст коефіцієнтів, побачити, наскільки сильно кожна ознака "відштовхує" межу розділення в той чи інший бік. Важливо також, що ми можемо оцінити, наскільки ефективний прогноз буде забезпечуватись з використанням різних значень порогової ймовірності класифікації.

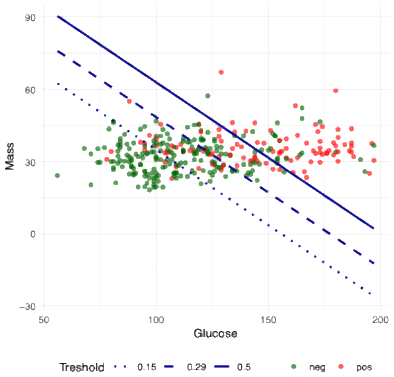


Рис. 8. Розподільчі межі для моделі логістичної регресії з двома предикторами glucose+mass, побудовані для порогових ймовірностей класифікації 0.15, 0.29 та 0.5

Для датасету 'PimaIndiansDiabetes2' такий графік наведено на Рис. 8. Зниження порогу класифікації зсуває лінію, яка є розподільчою межею для встановленого рівня ймовірності, вниз. Це дозволяє моделі бути більш чутливою (класифікувати більше випадків як pos), що критично важливо в медицині для мінімізації пропущених діагнозів (хибнонегативних результатів), навіть ціною збільшення кількості хибнопозитивних. Проте, як і у випадку з одним предиктором, модель не дуже добре розділяє класи. Для того, щоб діабет було виявлено майже у всіх, у кого він наявний, треба встановити дуже низький пороговий рівень ймовірності класифікації, наприклад, 0.15. І навіть за такого порогу будуть наявні хибнонегативні результати.

Важливо, що ми бачимо, що стандартний поріг класифікації на рівні 0.5 даватиме неприпустимо високий для медичної діагностики рівень хибнонегативних прогнозів, і для побудови більш ефективної моделі цей поріг необхідно буде суттєво знизити. Визначити оптимальний з певних міркувань поріг класифікації допомагають діаграми залежності таких

показників якості моделі як Sensitivity та Specificity від порогового рівня ймовірності та аналіз ROC-кривих після того, як будуть побудовані моделі на різних наборах предикторів, та оцінені метрики їх точності.

4. Метрики точності моделей логістичної регресії з різним набором вхідних ознак

Визначення метрик класифікації не є метою даної роботи, наголосимо лише на трьох з них, які є показовими та можуть бути критично важливими при прийнятті рішення щодо вибору моделі:

- загальна точність (Accuracy), яка показує загальну частку правильних прогнозів, найпростіша і найпоширеніша метрика, проте може бути оманливою на незбалансованих датасетах;

- повнота (Recall), вона ж – чутливість (Sensitivity) або рівень істинно позитивних прогнозів (True Positive Rate, TPR);

- специфічність (Specificity) або рівень істинно негативних прогнозів (True Negative Rate, TNR), та протилежна величина – рівень хибнопозитивних прогнозів (False Positive Rate, FPR), FPR = 1 – Specificity.

Вибір правильної метрики для оцінки моделі є критично важливим і залежить від конкретної задачі та наслідків різних типів помилок. Наприклад, в задачах діагностики певних хвороб критичною є наявність хибнонегативних результатів, коли особу з хворобою класифікують як здорову. В подібних випадках варто будувати модель з метою мінімізувати ймовірність хибнонегативних похибок. При цьому доведеться пожертвувати втратою точності класифікації здорових осіб: рівень хибнопозитивних класифікацій при цьому зросте, проте такі похибки часто є значно менш критичними.

Для аналізу та порівняння ми побудували різні моделі логістичної регресії для класифікації діабету: по всьому набору предикторів (Модель 1), з двома ознаками (рівень глюкози та ІМТ, Модель 2) та чотирма (рівень глюкози, ІМТ, вік та спадковість, Модель 3). Для порівняння ми додали модель наївного Баєсового класифікатора, побудованого по повному набору змінних (Модель 4). Базові метрики точності для цих моделей, отримані на тестовій частині датасету, наведено в Таблиці 1. Окрім вказаних вище метрик точності ми додали для моделей логістичної регресії (без деталізації, має бути мінімально можливим) також значення інформаційного критерію Акаїке – важливий показник якості самої побудованої моделі.

Таблиця 1. Порівняння метрик моделей на тестових даних: логістичної регресії за усіма наявними змінними (Модель 1), за змінними glucose+mass (Модель 2), glucose+mass+age+pedigree (Модель 3), модель наївного Баєсового класифікатора по повному набору змінних (Модель 4)

Метрика	Модель 1	Модель 2	Модель 3	Модель 4
Accuracy	0.8077	0.7821	0.8077	**0.8333**
Sensitivity (Recall)	0.4231	0.4615	0.4615	**0.6538**
Specifity	**1.000**	0.9423	0.9808	0.9231
AIC	306.05	310.44	300.83	

В даному випадку наївний баєсів класифікатор показує найкращі показники для ключових метрик. Він демонструє найвищу загальну точність (0.8333), тобто він робить більше правильних прогнозів в цілому. Попри те, що всі моделі логістичної регресії мають вищу специфічність, тобто трохи краще ідентифікують випадки відсутності діабету, наївний Баєсів класифікатор має значно вищу чутливість (0.6538 проти 0.4615). І в даному випадку це ключовий момент. Вища чутливість означає, що він значно краще розпізнає наявність діабету, роблячи менше помилкових негативних прогнозів. Це робить модель більш надійною для таких критичних завдань, як медична діагностика.

Тут ми не наводимо для порівняння точність моделей наївного баєсового класифікатора (НБК) з меншою кількістю предикторів, зауважимо лише, що для них результати гірші. Насправді це спостереження чудово ілюструє ключову відмінність між моделями логістичної регресії та наївного баєсового класифікатора. Головна причина полягає в тому, що НБК робить дуже сильне, в абсолютній більшості випадків несправедливе, але часто корисне припущення про незалежність предикторів. У ньому, навіть якщо змінна (triceps, pressure, insulin) не є ідеальним предиктором, вона все одно вносить свій маленький, але, як виявилося, позитивний внесок у фінальний прогноз. Кожна така змінна, навіть якщо вона здається "шумовою", додає ще один незалежний шматочок інформації, що в сукупності підвищує загальну якість моделі. Проте це має місце не для всіх датасетів.

Логістична регресія, навпаки, враховує всі предиктори разом, оцінюючи їхній сукупний вплив на результат. Якщо між предикторами є сильна залежність (наприклад, як між glucose та insulin), модель може "заплутатися" і надати меншої ваги одному з них, або ж його включення не принесе суттєвого покращення.

Порівнюючи метрики моделей логістичної регресі (Моделі 1-3) бачимо, що найбільша загальна точність досягається для повної моделі та моделі з чотирма предикторами. Проте повна модель, маючи найвищу специфічність, програє іншим в ключовому моменті: вона гірше виявляє діабет. І ми знову підкреслимо ключову проблему у виборі моделі: не існує однієї найкращої метрики. Вибір моделі завжди залежить від конкретної задачі та ціни помилок. В даному випадку з моделей логістичної регресії варто обрати одну зі спрощених моделей, спираючись на критичне для нашої задачі значення чутливості, і, окрім того, зважаючи на мінімальне значення AIC, ми оберемо і будемо намагатися покращити модель з чотирма предикторами: glucose + mass + age + pedigree.

Зауважимо, що не варто одразу відкидати модель логістичної регресії як таку, яка програє в точності моделі наївного баєсового класифікатора. Спочатку треба проаналізувати інші графіки та можливості підвищити критичні метрики точності обох моделей.

Очевидно, всі побудовані моделі далеко не бездоганні, значення чутливості все ж надто низьке. Проте їх можна покращити за рахунок зміни порогу класифікації, що ми зробимо, користуючись зображенням ROC-кривої та пов'язаних з нею графіків.

5. Оцінка ROC-кривої та оптимізація порогу класифікації

За замовчуванням, більшість класифікаторів, що видають ймовірність, використовують поріг 0.5. Якщо ймовірність належності до позитивного класу більша за 0.5, об'єкт класифікується як позитивний. Це працює добре, коли класи приблизно збалансовані, а вартість помилок (хибно-позитивних і хибно-негативних) однакова. Однак, у реальних задачах зазвичай класи не є збалансованими, один клас зустрічається набагато частіше за інший. Окрім того, в різних задачах ціна помилок дуже відрізняється. Вартість хибно-негативного прогнозу може бути значно вищою, ніж вартість хибно-позитивного і навпаки. Наприклад, у медичній діагностиці пропустити хворобу (хибно-негативний) набагато гірше, ніж поставити помилковий діагноз (хибно-позитивний), який потім буде спростовано додатковими обстеженнями.

Зміна стандартного порогу дозволяє змінити баланс між Sensitivity та Specificity відповідно до вимог конкретної задачі. Знижуючи поріг (наприклад, з 0.5 до 0.35), ми робимо модель більш чутливою до позитивного класу, збільшуючи Sensitivity, але водночас можемо збільшити кількість хибно-позитивних результатів (знизити Specificity). Існує декілька

поширених методів для визначення оптимального порогу. Одним з найбільш зручних та поширених є аналіз ROC-кривої (Receiver Operating Characteristic) та значення площі під нею, AUC (Area Under the ROC Curve), як допоміжної метрики для порівняння загальної здатності моделей розділяти класи.

У машинному навчанні ROC-крива – це графік, що показує співвідношення між True Positive Rate (чутливістю, Sensitivity) та False Positive Rate (1 – Specificity) для всіх можливих значень порогу класифікації (ймовірність приналежності об'єкту до цільового класу, яка використовується моделлю як межа) від 1 до 0. Вона допомагає оцінити здатність моделі розрізняти класи, незалежно від конкретного порогу.

Ідеальна точка на кривій знаходиться у лівому верхньому куті, де Sensitivity і Specificity є максимальними, що відповідає 100% чутливості та 0% хибнопозитивних спрацювань (TPR=1, FPR=0). Маючи зображення ROC-кривої, можна знайти поріг, який відповідає цій точці на кривій або іншій точці, яка відповідає оптимальним значенням TPR та FPR для конкретної задачі. Модель, яка класифікує випадковим чином, матиме ROC-криву, що проходить майже по діагоналі від (0,0) до (1,1).

Ще одним показником загальної якості моделі класифікатора є AUC – це площа під ROC-кривою. Значення AUC варіюється від 0 до 1. При цьому:

- AUC = 1 свідчить про ідеальну модель;
- AUC = 0.5 означає, що модель класифікує випадковим чином;
- AUC < 0.5 означає, що модель працює гірше за випадкову.

Отже, загалом високе значення AUC вказує на те, що модель добре розрізняє класи. AUC є особливо корисною метрикою для незбалансованих датасетів, оскільки вона не залежить від розподілу класів. На Рис. 9 можна побачити вигляд ROC-кривої для невдалої моделі класифікатора, де крива йде вздовж діагоналі, і значення AUC буде близьким до 0.5. Це свідчить про те, що рішення приймається абсолютно випадково, на основі такої інформації немає сенсу робити прогнози.

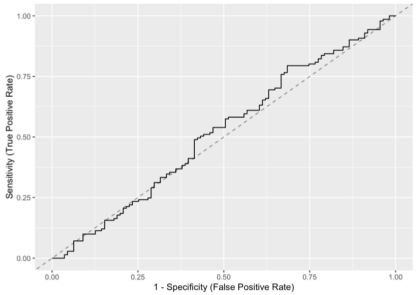


Рис. 9. Приклад ROC-кривої для невдалої моделі класифікації прогнозу динаміки курсу валют за даними кількох останніх днів

Оскільки загальна точність не завжди є найкращою метрикою, особливо для незбалансованих даних, ми можемо використати ROC-криву та AUC, як допоміжну метрику, для більш глибокого аналізу. Маючи графік ROC-кривої та значення AUC ми можемо порівняти загальну якість різних моделей (здатність до правильної класифікації), наприклад, логістичної регресії та НБК. Також, змінюючи поріг класифікації, ми можемо налаштовувати модель на краще виявлення того чи іншого класу. Графік ROC-кривої дозволяє побачити цю можливість наочно.

5.1. Оцінка ROC-кривої та оптимізація порогу класифікації для задачі виявлення діабету

В прикладі з класифікацією діабету нашою головною метою є зменшення ймовірності визнання хворої людини здоровою, тобто максимізація чутливості. Аналіз за допомогою ROC-кривих дозволить нам не лише порівняти загальну ефективність моделей, а й знайти ідеальний компроміс між чутливістю та специфічністю для кожної з них.

Ми будемо аналізувати дві моделі, які виявилися найкращими:

1) модель логістичної регресії, побудованої по чотирьох змінних glucose + mass + age + pedigree,

2) модель НБК, побудованого на всіх вхідних змінних.

Візуалізуємо ROC-криві для обох моделей на одному графіку (Рис. 10). Для того, щоб розуміти можливості моделі, на ROC-кривій помітимо точку,

яка відповідає порогу 0.5, який за замовчуванням використовується при класифікації об'єктів моделлю.

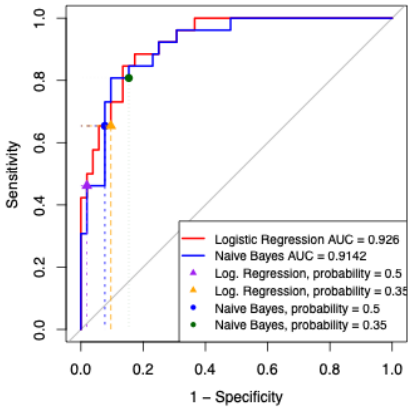


Рис. 10. ROC-криві для моделей класифікації наївної баєсівської та логістичної регресії з чотирма предикторами з різними пороговими ймовірностями

Треба звернути увагу на цьому зображенні на те, що обидві криві мають помітний вигин, отже, обидві моделі доволі непогані, і можуть забезпечити приблизно однакову точність класифікації при корекції порогу. І жодна з моделей не виглядає суттєво кращою за іншу. Також видно неозброєним оком, що стандартний поріг 0.5 дає далеко не найкращий результат. Зменшивши поріг, ми можемо суттєво підняти рівень чутливості для обох моделей. Отже, зміщення порогу в бік, наприклад, 0.35, є вдалою спробою підвищити чутливість (тобто, виявляти більше позитивних випадків) за рахунок специфічності (модель частіше помилково ідентифікуватиме негативні випадки як позитивні). Як бачимо, в обох моделях чутливість суттєво зросла (точки посунулись вище), при невеликих втратах в значеннях специфічності (точки посунулись праворуч), що і було нашою метою. Проте, очевидно, ми можемо ще опустити поріг класифікації для покращення результатів.

Графік на Рис. 10 демонструє, що зміна порогу для кожної з моделей може суттєво вплинути на ефективність класифікації. Для визначення оптимального порогу є різні методи, які спираються на різні критерії

(наприклад, визначення найближчої точки до верхнього лівого кута ROC-кривої або максимізації певної метрики, як-от F1-score). Знову таки, поняття оптимального порогу залежить від конкретної задачі. Для того, щоб побачити динаміку метрик якості, побудуємо окремо графіки залежності чутливості, специфічності та середньої точності від значення порогової ймовірності та спробуємо побачити прийнятне для нашої задачі її значення (Рис. 11).

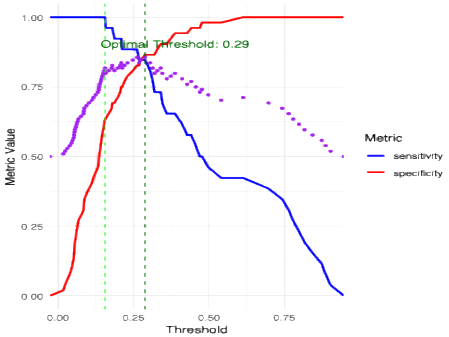


Рис. 11. Залежність базових метрик (чутливості, специфічності та збалансованої точності) від порогової ймовірності класифікації для моделі класифікації логістичної регресії з чотирма предикторами. Пунктирними лініями позначені порогові ймовірності на рівні 0.155 (світло-зелена) та 0.29 (темно-зелена).

Бачимо, що стовідсоткового виявлення діабету можна досягнути, якщо встановити пороговий рівень ймовірності класифікації 0.155 (світло-зелена пунктирна лінія). При цьому рівень специфічності буде приблизно 0.63. Насправді, зважаючи на те, що ця похибка є суттєво менш критичною, це цілком придатний показник. Якщо йти шляхом максимізації збалансованої точності, то варто обрати рівень порогової ймовірності 0.29.

Метрики точності цієї моделі (логістична регресія, 4 предиктори, пороговий рівень ймовірності класифікації 0.29) зображені на Рис. 12. Бачимо, що в результаті зміни порогової ймовірності класифікації ми отримали не лише доволі високий рівень чутливості, але й збільшення загальної точності класифікації. При цьому втрата в значенні специфічності несуттєва. Знову таки, якщо стоїть задача мінімізувати ймовірність хибнонегативного результата, можна, користуючись останніми діаграмами, ще опустити рівень ймовірності класифікації для підвищення чутливості.

```
> print(conf_matrix_logreg_new_thresh)
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg  45   4
       pos   7  22

               Accuracy : 0.859
                 95% CI : (0.7617, 0.9274)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.0001012

                  Kappa : 0.6916

 Mcnemar's Test P-Value : 0.5464936

            Sensitivity : 0.8462
            Specificity : 0.8654
         Pos Pred Value : 0.7586
         Neg Pred Value : 0.9184
             Prevalence : 0.3333
         Detection Rate : 0.2821
   Detection Prevalence : 0.3718
      Balanced Accuracy : 0.8558

       'Positive' Class : pos
```

Рис. 12. Матриця похибок та інші метрики для моделі логістичної регресії виявлення діабету за чотирма змінними з порогом класифікації 0.29

Подивимось на аналогічні зображення для наївного баєсового класифікатора на Рис. 13. Бачимо, що модель доволі суттєво відрізняється від моделі логістичної регресії за динамікою метрик залежно від порогової ймовірності. Зокрема 100% випадків діабету на тестовій вибірці модель виявляє при пороговому рівні 0.022, при цьому модель буде давати хибнопозитивний прогноз майже в половині випадків. Це значно гірша ситуація, ніж в попередній моделі.

Максимальне значення збалансованої точності визначено на рівні порогу класифікації 0.38. Якщо встановити цей поріг в модель, матимемо результати для метрик, наведені на Рис. 14. Бачимо високу загальну точність, але рівень чутливості дещо нижчий, ніж в моделі логістичної регресії.

На наш погляд, знову таки, маючи на меті підвищення чутливості без великих втрат в специфічності, можемо для баєсової моделі ще трохи зменшити пороговий рівень класифікації. На Рис. 13 видно, що якщо зсунути його до рівня 0.305, ми дещо виграємо в чутливості і небагато втратимо в специфічності. Якщо встановити цей поріг в моделі наївного баєсового

класифікатора, ми дещо втратимо в загальній точності, проте досягнемо рівня чутливості як в моделі логістичної регресії з порогом 0.29.

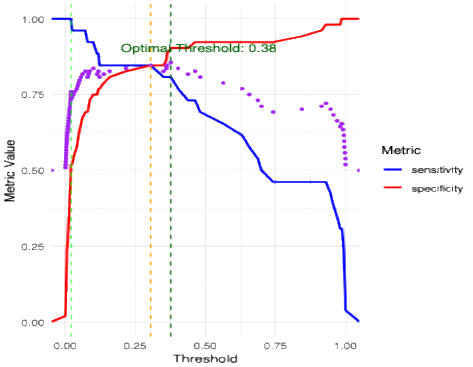


Рис. 13. Залежність базових метрик (чутливості, специфічності та збалансованої точності) від порогової ймовірності класифікації для наївної баєсової моделі класифікації. Пунктирними лініями позначені порогові ймовірності 0.022 (світло-зелена), 0.305 (помаранчева) та 0.38 (темно-зелена)

```
Confusion Matrix and Statistics

          Reference
Prediction neg pos
       neg  47   5
       pos   5  21

               Accuracy : 0.8718
                 95% CI : (0.7768, 0.9368)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 3.132e-05

                  Kappa : 0.7115

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8077
            Specificity : 0.9038
         Pos Pred Value : 0.8077
         Neg Pred Value : 0.9038
             Prevalence : 0.3333
         Detection Rate : 0.2692
   Detection Prevalence : 0.3333
      Balanced Accuracy : 0.8558

       'Positive' Class : pos
```

Рис. 14. Матриця похибок та інші метрики для наївної баєсової моделі виявлення діабету з пороговим рівнем класифікації 0.38.

Отже, для нашого датасету завдяки зсуву порогового рівня класифікації вдалось дуже суттєво покращити якість обох моделей, що було б важче зробити без відповідних діаграм. При цьому модель логістичної регресії при однаково високій чутливості виглядає дещо кращою за НБК за рахунок вищої специфічності (0.865 проти 0.846). Проте в таких випадках при порівнянні моделей обов'язково треба дати відповідь на питання: чи не є це випадковістю?

6. Статистична перевірка ідентичності моделей

При виборі оптимальної моделі треба усвідомлювати, що отримані метрики точності (зокрема Accuracy, AUC, Sensitivity) є точковими оцінками, обчисленими для конкретної реалізації випадкової вибірки та специфічного її розбиття на тренувальну та тестову частини. Оскільки ми оперуємо вибірковими даними, оцінка будь-якої метрики є випадковою величиною, значення якої залежить від варіативності вхідних даних. Отже, перевага однієї моделі над іншою може бути наслідком випадкового збігу, а не реальної різниці в узагальнювальній здатності. Якщо розбити вибірку на тренувальну та тестову частини в інший спосіб, метрики можуть змінитися, і найкраща модель серед кількох альтернативних може втратити лідерську позицію. Очевидно, якщо взяти інший набір даних, то також результати точності можуть бути зовсім іншими.

Для прийняття обґрунтованого рішення щодо вибору фінальної архітектури класифікатора необхідно провести формальний статистичний аналіз із використанням наступного інструментарію:

1. Аналіз довірчих інтервалів (CI) для Accuracy (виводяться в стандартному звіті метрик моделі, див. Рис. 12, 14). Ми розглядаємо не лише точкову оцінку, а й інтервал, у якому з заданою надійністю (зазвичай 0.95) знаходиться істинне значення параметру. Якщо довірчі інтервали двох моделей мають значну зону перекриття, ми не маємо статистичних підстав стверджувати про перевагу однієї з них, навіть якщо її середня точність вища. Це свідчить про високу дисперсію помилки моделі.

2. Критерій ДеЛонга (DeLong's Test) для порівняння AUC – загальної здатності моделей розділяти класи. Оскільки AUC є інтегральною метрикою якості розділення класів, ми використовуємо критерій ДеЛонга для перевірки нульової гіпотези $H_0: AUC_1 = AUC_2$. Цей тест враховує кореляцію між кривими (оскільки вони побудовані на одній вибірці) і дозволяє визначити,

чи є різниця в роздільній здатності моделей статистично значущою ($p - value < 0.05$).

3. Критерій Мак-Немара (McNemar's Test). Цей непараметричний тест є критично важливим при зміні порогів класифікації. Він аналізує узгодженість помилок моделей, фокусуючись на неузгоджених парах (випадках, де одна модель помилилася, а інша – ні). Нульова гіпотеза стверджує, що ймовірності помилок обох моделей однакові. Це дозволяє нам статистично підтвердити, що зсув порогу дійсно приводить до значущої зміни структури помилок (зменшення, наприклад, ймовірності похибки II роду за рахунок похибки I роду).

Не наводячи деталей, зауважимо, що статистичні тести для задачі виявлення діабету показують загальну ідентичність роботи цих двох обраних моделей, тобто формально немає математичної підстави вважати одну модель кращою за іншу на цій вибірці. При скоригованих порогах моделей ЛР та НБК критерій Мак-Немара дає p-значення, рівне 0.01333, тобто на рівні значущості 0.05 ми маємо відхилити нульову гіпотезу про рівність моделей. Проте воно не є переконливим і до висновку треба ставитися з обережністю.

У випадку, коли моделі статистично рівні, варто обрати ту, яка простіша в інтерпретації (зазвичай це логістична регресія) або ту, яка має менше "дорогих" помилок (False Negatives) в конкретній задачі.

7. Висновки.

У роботі обґрунтовано концептуальну значущість міждисциплінарної інтеграції методів візуалізації на всіх етапах життєвого циклу побудови класифікаційних моделей. Проведене дослідження дозволяє зробити наступні узагальнюючі висновки:

1. Роль візуальної аналітики на етапі препроцесингу: Застосування комплексних матричних діаграм (Matrix Plots) та аналізу топології даних у просторі головних компонент (PCA) дозволяє не лише виявити приховані закономірності та колінеарні зв’язки (зокрема, між віком пацієнта та репродуктивним анамнезом), а й ідентифікувати критичні аномалії в даних. Це підтверджує, що візуалізація є першим і необхідним етапом верифікації якості вибірки. Без візуального контролю математичні алгоритми ризикують оперувати артефактами даних (маскованими пропусками чи нетиповими викидами), що нівелює подальшу точність моделювання.

2. Геометрична інтерпретація статистичних рішень: Встановлено, що перехід від багатовимірних матриць до аналізу проєкцій на головні площини та візуалізації меж прийняття рішень (Decision Boundaries) дозволяє

дестилювати "чорну скриньку" алгоритмів машинного навчання. Візуальна оцінка зон перекриття класів дає досліднику інтуїтивне розуміння фізичного змісту моделі, яке неможливо отримати виключно через інтегральні метрики точності.

3. Оптимізація структури моделі: Шляхом порівняльного аналізу конфігурацій логістичної регресії встановлено, що нарощування кількості предикторів не завжди призводить до покращення прогнозної сили. Використання інформаційного критерію Акаїке (AIC) спільно з аналізом факторних навантажень PCA дозволило обґрунтувати вибір оптимальної чотирьохфакторної моделі, яка забезпечує найкращу узагальнювальну здатність.

4. Візуальне обґрунтування вибору порогів: Продемонстровано, що візуалізація логіт-трансформацій та ймовірнісних кривих є фундаментальним підґрунтям для стратегічної зміни порогів класифікації. Це дозволяє усвідомлено керувати балансом між помилками першого та другого роду, що є критично важливим у прикладних задачах, де ціна пропуску цільової події є надзвичайно високою.

5. Методологічна цілісність: Запропонований підхід доводить, що статистична верифікація (критерії ДеЛонга, Мак-Немара) та графічна аналітика є взаємодоповнюючими компонентами. Візуалізація формує гіпотезу та стратегію моделювання, а формальні критерії забезпечують математичну строгість отриманих результатів.

Отже, інтеграція методів візуалізації на кожному етапі машинного навчання перетворює процес побудови моделі з «чорної скриньки» на прозорий, верифікований інструмент підтримки прийняття рішень. Такий підхід мінімізує ризики формального застосування алгоритмів і дозволяє створювати моделі, що мають не лише високі математичні метрики, а й реальну діагностичну цінність.

Список використаних джерел

1. Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2026, pp. 1135-1144: https://arxiv.org/abs/1602.04938

2. Lundberg, S. M., & Lee, S. I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, №30, 2017: https://arxiv.org/abs/1705.07874

3. Shneiderman, Ben. The Big Picture for Big Data: Visualization. Science (New York). 2014. 343. 730. 10.1126/science.343.6172.730-a.

4. Hanna Livinska, Daria Skrypnyk, José Luis GALÁN-GARCÍA: Comparative Study of Machine Learning Algorithms for Diabetes Detection Using Binary Data, *"Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics"*, №1(78)/2024 -- Ст. 119-127.

5. Blake, C.L. & Merz, C.J. UCI Repository of Machine Learning Databases. Irvine, CA: University of California, Irvine, Department of Information and Computer Science. Formerly available from http://www.ics.uci.edu/~mlearn/MLRepository.html.