УДК 004.67 + 616.2

Р. А. Бралатан
С. О. Жуков

ПОРІВНЯННЯ NAIVE BAYES ТА TREE-AUGMENTED NAIVE BAYES ДЛЯ ПЕРЕДБАЧЕННЯ РИЗИКУ ЦУКРОВОГО ДІАБЕТУ

Вінницький національний технічний університет, м. Вінниця, Україна

Анотація

У роботі досліджено можливість удосконалення класичного байєсівського класифікатора Naive Bayes шляхом введення деревоподібних залежностей між ознаками у моделі Tree-Augmented Naive Bayes (TAN). Експеримент проведено на збалансованій вибірці BRFSS 2015 із 70 692 спостережень. Безперервні предиктори дискретизовано на квантильні інтервали, а структуру залежностей TAN побудовано методом максимального кістякового дерева на основі умовної взаємної інформації між ознаками. Обидві моделі навчено зі згладжуванням Лапласа та порівняно за показниками Accuracy, Precision, Recall, F1-score, AUC-ROC і Brier score, що дозволило одночасно оцінити їхню дискримінаційну здатність і якість калібрування ймовірностей. Модель TAN досягла Accuracy 0.7458, F1-score 0.7555 та AUC-ROC 0.8222, що перевищує відповідні показники базового Naive Bayes (Accuracy 0.7325, F1-score 0.7311, AUC-ROC 0.8123), а також забезпечила менший Brier score (0.1775 проти 0.2000). Додатково проаналізовано найсильніші міжознакові залежності, виявлені у структурі TAN. Встановлено, що врахування умовних залежностей між клінічними та соціально-поведінковими предикторами підвищує якість байєсівської класифікації ризику діабету.

Ключові слова: байєсівська мережа, Naive Bayes, Tree-Augmented Naive Bayes, цукровий діабет, прогнозування ризику, збалансовані дані, калібрування.

Abstract

This paper investigates the improvement of the classical Naive Bayes classifier through the introduction of tree-structured feature dependencies in Tree-Augmented Naive Bayes (TAN). The experiment was conducted on a balanced BRFSS 2015 sample of 70,692 observations. Continuous predictors were discretized into quantile intervals, and the TAN dependency structure was built using the maximum spanning tree method based on the conditional mutual information between features. Both models were trained with Laplace smoothing and compared in terms of Accuracy, Precision, Recall, F1-score, AUC-ROC, and Brier score, which allowed simultaneous assessment of their discriminative ability and the quality of probability calibration. The TAN model achieved Accuracy 0.7458, F1-score 0.7555, and AUC-ROC 0.8222, outperforming the baseline Naive Bayes (Accuracy 0.7325, F1-score 0.7311, AUC-ROC 0.8123), and also yielded a lower Brier score (0.1775 versus 0.2000). In addition, the strongest inter-feature dependencies identified in the TAN structure were analyzed. The study demonstrates that modeling conditional dependencies among clinical and socio-behavioral predictors enhances Bayesian diabetes risk classification.

Keywords: Bayesian network, Naive Bayes, Tree-Augmented Naive Bayes, diabetes, risk prediction, balanced data, calibration.

Вступ

Проблема раннього прогнозування ризику цукрового діабету є актуальною для сучасних систем охорони здоров'я, оскільки своєчасне виявлення захворювання забезпечує обґрунтованість превентивних рішень та раціональне використання клінічних ресурсів [1, 2].

Байєсівські моделі у цій задачі є методологічно доцільними через їхню здатність представляти невизначеність у формі ймовірностей і водночас зберігати інтерпретованість структурних залежностей між медичними предикторами [2]. Використання збалансованих даних для порівняння моделей дозволяє отримати коректні та незміщені оцінки якості класифікації [1].

Метою роботи є порівняльне оцінювання класичного Naive Bayes та Tree-Augmented Naive Bayes для задачі прогнозування ризику цукрового діабету на збалансованій вибірці BRFSS 2015 і кількісне визначення впливу врахування міжознакових залежностей на якість класифікації та калібрування ймовірностей. Для досягнення поставленої мети розв'язано такі завдання: підготовлено та дискретизовано

дані, побудовано структуру залежностей TAN, навчено обидві моделі та виконано їх порівняння за сукупністю метрик якості й калібрування.

Методологія

Класичний Naive Bayes ґрунтується на припущенні умовної незалежності предикторів за відомого класу, що спрощує оцінювання параметрів, однак може обмежувати якість моделювання в умовах реальних клінічних взаємозв'язків між ознаками [2, 7].

Tree-Augmented Naive Bayes усуває зазначене обмеження шляхом введення деревоподібних залежностей між ознаками, зберігаючи при цьому прозорість імовірнісної інтерпретації, що відповідає вимогам до клінічних інструментів підтримки рішень [6]. Удосконалення базової моделі за рахунок ускладнення її структури є поширеною стратегією підвищення прогностичної точності, як і в ансамблевих методах прогнозування, зокрема бустингу [3]. Попередні роботи підтверджують, що байєсівські мережі є практичним інструментом для медичної діагностики та підтримки клінічних рішень [1, 6].

Апріорний розподіл класу моделі задається співвідношенням, поданим як формула (1):

$$P(y = c) = \pi_c, \tag{1}$$

де y — класова змінна, c — конкретний клас, π_c — апріорна ймовірність класу.

Функція правдоподібності для моделі Naive Bayes визначається як формула (2):

$$P(x|y = c) = \prod(j = 1..d)\ P(x_j|y = c), \tag{2}$$

де x — вектор ознак, d — кількість ознак, x_j — j-та ознака, $P(x_j|y = c)$ — умовна ймовірність j-тої ознаки за відомого класу.

Для Tree-Augmented Naive Bayes правдоподібність обчислюється як формула (3):

$$P(x|y = c) = P(x_r|y = c) \cdot \prod(j \neq r)P\left(x_j|x_{pa}(j), y = c\right), \tag{3}$$

де x_r — коренева ознака дерева, $x_{pa}(j)$ — батьківська ознака j-тої вершини, умовна ймовірність враховує клас і міжознакову залежність.

Апостеріорна ймовірність за теоремою Байєса обчислюється як формула (4):

$$P(y = c|x) = P(y = c)P(x|y = c)/\Sigma_k\ P(y = k)P(x|y = k), \tag{4}$$

де C — множина класів, P(y=c|x) — апостеріорна ймовірність класу c для спостереження x, знаменник виконує нормування за всіма класами.

Цільова метрика узгодження точності і повноти визначається через формулу (5):

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall}), \tag{5}$$

де Precision — точність позитивних класифікацій, Recall — повнота виявлення позитивного класу, F1 — гармонійне середнє зазначених показників [5, 9].

Матеріали та методи

Експеримент проведено на наборі diabetes_binary_5050split_health_indicators_BRFSS2015.csv із обсягом 70692 спостережень і балансом класів 50/50 (35 346 записів кожного класу), що дозволяє уникнути систематичного перекосу на користь одного класу під час порівняння моделей [3, 4].

Безперервні ознаки BMI, MentHlth та PhysHlth дискретизовано на 5 квантильних інтервалів. Усі категоріальні змінні представлено цілочисельними кодами з ненегативними значеннями, що відповідає вимогам моделі CategoricalNB зі згладжуванням Лапласа (α = 1.0) [7].

Порівняння виконано для базового Naive Bayes та Tree-Augmented Naive Bayes за показниками Accuracy, Precision, Recall, F1-score, AUC-ROC і Brier score для сумісної оцінки дискримінаційних властивостей і калібрування імовірностей [9].

Баланс класів у вибірці представлено на рисунку 1.

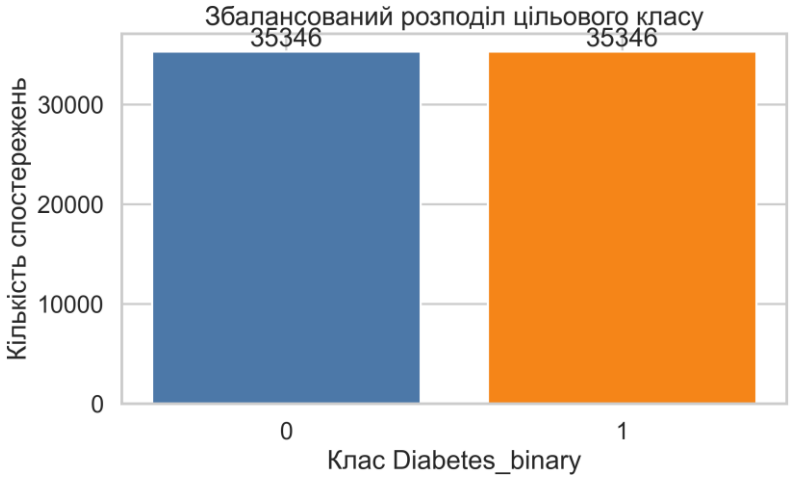

Рис. 1. Збалансований розподіл цільового класу

Як видно з рисунка 1, обидва класи цільової змінної Diabetes_binary представлено однаковою кількістю спостережень — по 35 346 записів. Така рівність часток усуває апріорний нахил класифікатора на користь мажоритарного класу і гарантує, що відмінності у значеннях метрик відображають саме властивості моделей, а не дисбаланс вибірки. Це створює коректну основу для подальшого порівняння Naive Bayes і TAN.

Структурні залежності в моделі TAN, отримані методом максимального кістякового дерева, подано на рисунку 2.

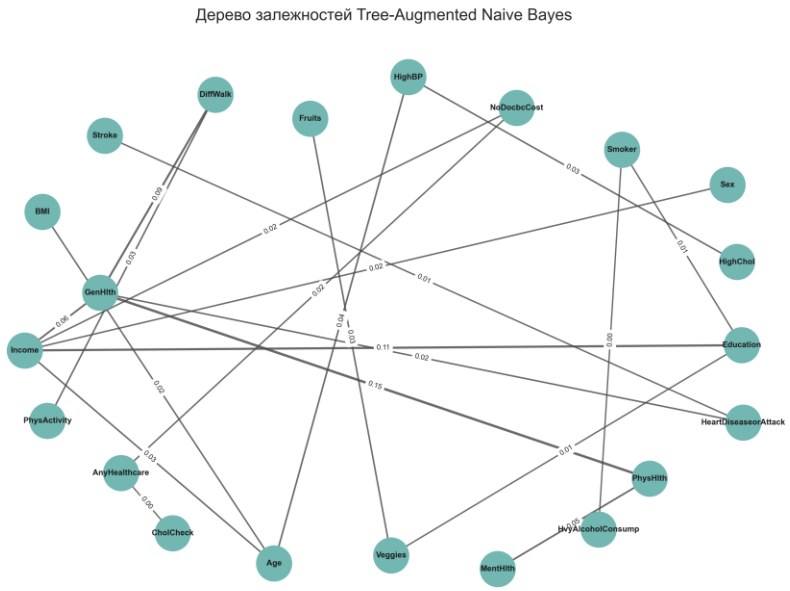

Рис. 2. Дерево залежностей Tree-Augmented Naive Bayes

Рисунок 2 ілюструє деревоподібну структуру, у якій кожна ознака, окрім кореневої, отримує одну батьківську вершину, а вага ребра відповідає умовній взаємній інформації між ознаками. Найтовстіші зв'язки об'єднують показники загального та фізичного стану здоров'я (GenHlth, PhysHlth, DiffWalk) і соціально-економічні предиктори (Education, Income), що узгоджується з клінічними уявленнями про

взаємозалежність цих чинників. Саме ці залежності, які класичний Naive Bayes ігнорує через припущення про незалежність ознак, і використовує модель TAN для уточнення оцінок ймовірностей.

Результати дослідження

В таблиці 1 наведені основні метрики порівняння моделей.

Таблиця 1. Основні метрики порівняння моделей

Модель	Accuracy	Precision	Recall	F1-score	AUC-ROC	Brier score
Naive Bayes	0.7325	0.7350	0.7273	0.7311	0.8123	0.2000
Tree-Augmented Naive Bayes	0.7458	0.7277	0.7855	0.7555	0.8222	0.1775

Naive Bayes досягла Accuracy 0.7325, Precision 0.7350, Recall 0.7273, F1-score 0.7311, AUC-ROC 0.8123 та Brier score 0.2000. Tree-Augmented Naive Bayes забезпечила Accuracy 0.7458, Precision 0.7277, Recall 0.7855, F1-score 0.7555, AUC-ROC 0.8222 і Brier score 0.1775 [4]. В таблиці 2 наведено порівняння основних метрик для TAN та базового Naive Bayes. В таблиці 3 наведені найсильніші залежності у структурі TAN.

Таблиця 2. Порівняння TAN та базового Naive Bayes

Показник	Зміна
Accuracy	+0.0133
Precision	-0.0073
Recall	+0.0583
F1-score	+0.0244
AUC-ROC	+0.0099
Brier score	-0.0225

Таблиця 3. Найсильніші залежності у структурі TAN

Ознака 1	Ознака 2	Conditional Mutual Information
GenHlth	PhysHlth	0.1531
Education	Income	0.1127
GenHlth	DiffWalk	0.0937
GenHlth	Income	0.0598
MentHlth	PhysHlth	0.0541
HighBP	Age	0.0383
Age	Income	0.0308
PhysActivity	DiffWalk	0.0277

Порівняння ROC-кривих обох моделей наведено на рисунку 3.

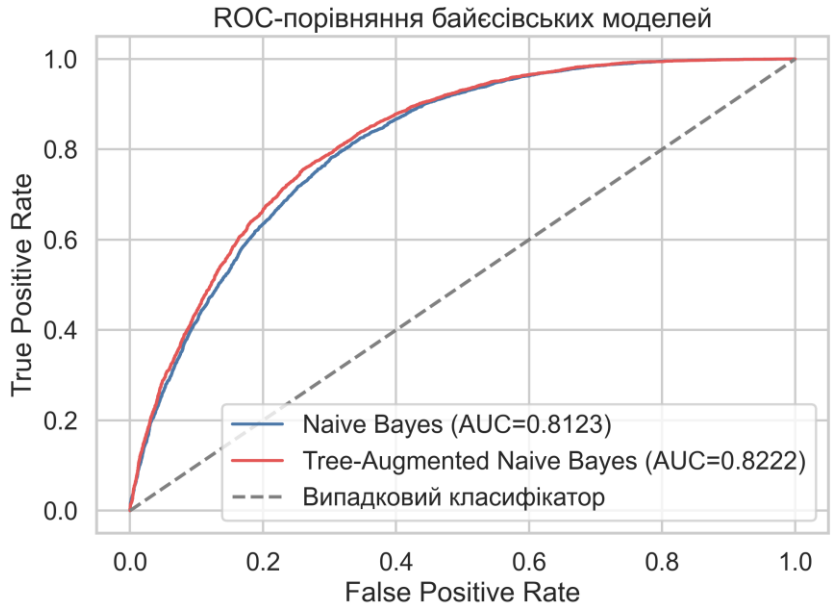

Рис. 3. Порівняння ROC-кривих

На рисунку 3 крива TAN розташована вище від кривої базового Naive Bayes майже на всьому діапазоні значень False Positive Rate, що відповідає більшій площі під кривою (AUC = 0.8222 проти 0.8123). Це означає, що за фіксованої частки хибнопозитивних висновків модель TAN виявляє більше дійсних випадків ризику діабету. Обидві криві помітно віддалені від діагоналі випадкового класифікатора, що свідчить про практично значущу дискримінаційну здатність байєсівського підходу загалом.

Порівняння калібрування імовірностей наведено на рисунку 4.

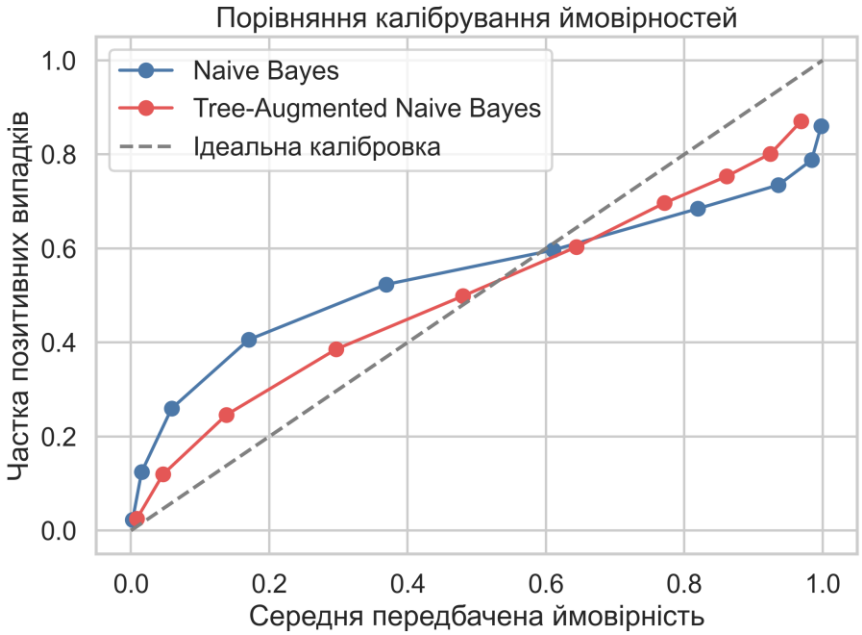

Рис. 4. Порівняння калібрування ймовірностей

Рисунок 4 показує, що крива калібрування TAN розташована ближче до лінії ідеальної калібровки, ніж крива Naive Bayes, особливо в області низьких прогнозованих ймовірностей, де базова модель помітно завищує оцінку ризику. Краще калібрування TAN підтверджується і нижчим значенням Brier score (0.1775 проти 0.2000). Це важливо для медичних застосувань, оскільки рішення про скринінг спираються не лише на бінарну мітку, а й на узгодженість самого значення ймовірності ризику з фактичною частотою захворювання.

Сукупність метрик свідчить про статистично стабільну перевагу TAN у якості класифікації та надійності імовірнісних оцінок, що відповідає очікуваному ефекту врахування залежностей між предикторами [6, 10].

Висновки

У роботі виконано порівняння класичного Naive Bayes і Tree-Augmented Naive Bayes для задачі передбачення ризику цукрового діабету на збалансованій вибірці BRFSS 2015. Перевага TAN над базовою Naive Bayes підтверджена метриками: приріст Accuracy становить +0.0133, приріст Recall +0.0583, приріст F1-score +0.0244, приріст AUC-ROC +0.0099, тоді як Brier score зменшується на 0.0225. Показники AUC-ROC = 0.8222 (TAN) та AUC-ROC = 0.8123 (Naive Bayes) свідчать про практично значущу дискримінаційну здатність обох моделей, при цьому TAN демонструє кращу узгодженість ймовірностей з фактичними наслідками, що особливо важливо для медичних систем підтримки рішень.

У підсумку встановлено, що Tree-Augmented Naive Bayes є доцільнішою за базову Naive Bayes для задачі передбачення ризику цукрового діабету на дослідженому наборі даних, оскільки поєднує інтерпретованість байєсівської моделі з додатковою прогностичною потужністю, яка досягається за рахунок урахування релевантних міжознакових залежностей. Найсильніші залежності між GenHlth та PhysHlth (CMI = 0.1531), Education та Income (CMI = 0.1127) і GenHlth та DiffWalk (CMI = 0.0937) підтверджують, що стан здоров'я, освіта і фізичний стан є взаємопов'язаними вимірами ризику, моделювання яких покращує прогностичну точність.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Rodinkova V., Yuriev S., Mokin V., Kryvopustova M., Shmundiak D., Bortnyk M., Kryzhanovskyi Y., Kurchenko A. Bayesian analysis suggests independent development of sensitization to different fungal allergens. World Allergy Organization Journal. 2024. Vol. 17, no. 5. P. 100908. [Електронний ресурс] – Режим доступу: https://doi.org/10.1016/j.waojou.2024.100908
2. Бралатан Р. А., Жуков С. О. Байєсівське моделювання для оцінювання ризиків виникнення раку легенів на основі аналізу медичних даних. Матеріали LIV Всеукраїнської науково-технічної конференції підрозділів Вінницького національного технічного університету (НТКП ВНТУ–2025). Вінниця, 2025. [Електронний ресурс] – Режим доступу: https://press.vntu.edu.ua/index.php/vntu/catalog/book/904
3. Копняк В. Є., Мокін В. Б., Жуков С. О., Варчук І. В., Скринник Т. В. Метод бустингу гетероскедастичних моделей для прогнозування концентрацій пилу Сахари в атмосферному повітрі України. Наукові праці Вінницького національного технічного університету. 2024. № 2. [Електронний ресурс] – Режим доступу: https://doi.org/10.31649/2307-5376-2024-2-28-38
4. Zhao Z. et al. Bayesian Cox regression for large-scale inference with applications to electronic health records. The Annals of Applied Statistics. 2023. [Електронний ресурс] – Режим доступу: https://doi.org/10.1214/22-AOAS1658
5. Choi B. G. et al. Predicting Current Glycated Hemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm. JMIR Medical Informatics. 2020. Vol. 8, No. 9. [Електронний ресурс] – Режим доступу: https://doi.org/10.2196/18963
6. Lu Y. et al. Medical idioms for clinical Bayesian network development. Journal of Biomedical Informatics. 2020. Vol. 110. Art. 103495. [Електронний ресурс] – Режим доступу: https://doi.org/10.1016/j.jbi.2020.103495
7. Kourou K. et al. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015. Vol. 13. P. 8–17. [Електронний ресурс] – Режим доступу: https://doi.org/10.1016/j.csbj.2014.11.005
8. Swanson K., Wu E. Q., Zhang A., Alizadeh A. A., Zou J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023. Vol. 186, No. 8. P. 1772–1791. [Електронний ресурс] – Режим доступу: https://doi.org/10.1016/j.cell.2023.01.035
9. Topol E. J. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019. Vol. 25. P. 44–56. [Електронний ресурс] – Режим доступу: https://doi.org/10.1038/s41591-018-0300-7
10. Roman Bralatan notebook [Електронний ресурс] – Режим доступу: https://www.kaggle.com/code/romantick/nb-vs-tan-diabetes-risk-bayesian-comparison

Бралатан Роман Андрійович – студент групи 124-24а, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: bralatan.roman@gmail.com

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@vntu.edu.ua

Bralatan Roman A. - student of Faculty of Intelligent Information Technologies and Automation, 124-24a, Vinnytsia National Technical University, Vinnytsia, e-mail bralatan.roman@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@vntu.edu.ua