УДК 004.852:616.379

М. І. Янковчук
С. О. Жуков

СИСТЕМНИЙ АНАЛІЗ ТА ІМОВІРНІСНЕ МОДЕЛЮВАННЯ РИЗИКУ ВИНИКНЕННЯ ДІАБЕТУ НА ОСНОВІ БАЙЄСІВСЬКИХ МЕТОДІВ

Вінницький національний технічний університет

Анотація

У роботі проведено системний аналіз застосування імовірнісних методів класифікації для оцінки ризику виникнення цукрового діабету на основі набору даних Pima Indians Diabetes. Виконано комплексне очищення та попередню обробку даних, побудовано та досліджено моделі GaussianNB і CategoricalNB із застосуванням різних стратегій дискретизації. Проведено порівняльний аналіз ефективності логарифмічного перетворення, оптимізації гіперпараметрів та ансамблевих підходів. Встановлено, що найвищу точність серед імовірнісних моделей демонструє GaussianNB з адаптивним згладжуванням, тоді як використання ансамблевих методів (Voting Ensemble) дозволяє максимізувати показник ROC-AUC для систем підтримки прийняття медичних рішень.

Ключові слова: системний аналіз, імовірнісне моделювання, GaussianNB, CategoricalNB, дискретизація даних, прогнозування діабету, машинне навчання, предиктивна діагностика.

Abstract

The paper conducts a systems analysis of probabilistic classification methods for assessing diabetes mellitus risk based on the Pima Indians Diabetes dataset. Comprehensive data cleaning and preprocessing were performed, and GaussianNB and CategoricalNB models were constructed and investigated using various discretization strategies. A comparative analysis of logarithmic transformation efficiency, hyperparameter optimization, and ensemble approaches was carried out. It was established that GaussianNB with adaptive smoothing demonstrates the highest accuracy among probabilistic models, while the use of ensemble methods (Voting Ensemble) allows for maximizing the ROC-AUC indicator for medical decision support systems.

Keywords: systems analysis, probabilistic modeling, GaussianNB, CategoricalNB, data discretization, diabetes prediction, machine learning, predictive diagnostics.

Вступ

Цукровий діабет є одним із найпоширеніших хронічних захворювань у світі, рання діагностика якого суттєво знижує ризик ускладнень. Застосування методів машинного навчання для автоматизованого прогнозування ризику захворювання дозволяє підтримувати прийняття клінічних рішень на основі об'єктивних даних.

Байєсівські класифікатори є особливо привабливими для медичних застосувань завдяки своїй інтерпретованості, обчислювальній ефективності та природній здатності працювати з імовірнісними оцінками. Теорема Байєса дозволяє обчислити апостеріорну ймовірність належності пацієнта до класу «діабет» на основі спостережуваних клінічних показників.

Метою роботи є побудова та порівняльний аналіз байєсівських моделей класифікації для прогнозування ризику діабету, дослідження впливу попередньої обробки даних і дискретизації на якість прогнозів.

Попередня обробка та аналіз даних

Для дослідження використано набір даних Pima Indians Diabetes Dataset [1], який містить 768 записів із 8 клінічними ознаками: кількість вагітностей, рівень глюкози, артеріальний тиск, товщина шкірної складки, рівень інсуліну, індекс маси тіла (BMI), функція діабетичної спадковості та вік. Цільова змінна — наявність діабету (0 або 1). Розподіл класів є незбалансованим: 65,1% негативних і 34,9% позитивних випадків.

На етапі очищення даних [2] виявлено, що нульові значення в медично неможливих полях фактично є пропущеними: Glucose — 5 записів, BloodPressure — 35, SkinThickness — 227, Insulin — 374, BMI — 11. Такі значення замінено на NaN. Після видалення рядків, де кількість пропусків перевищувала допустиму межу, обсяг вибірки склав 724 записи.

Аналіз розподілів ключових ознак (рис. 1) показав, що жодна з них не відповідає нормальному закону: тест Шапіро–Вілка дав p-value < 0.05 для Glucose, BMI та Insulin. Графіки розподілів ознак зображені нижче (рис. 1).

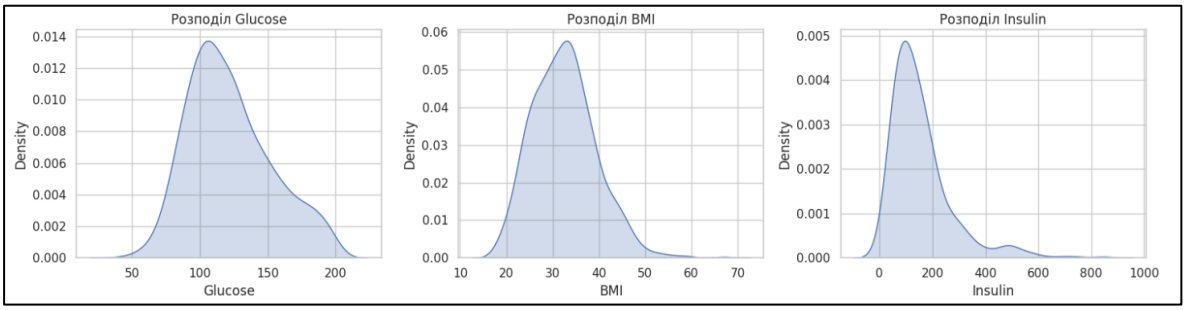

Рис. 1. KDE-графіки розподілів Glucose, BMI, Insulin

На рис. 1 видно, розподіл Insulin є найбільш скошеним із правостороннім хвостом, що підтверджує значну варіативність цього показника серед пацієнтів..

Для відбору інформативних ознак побудовано теплову мапу кореляцій (рис. 2).

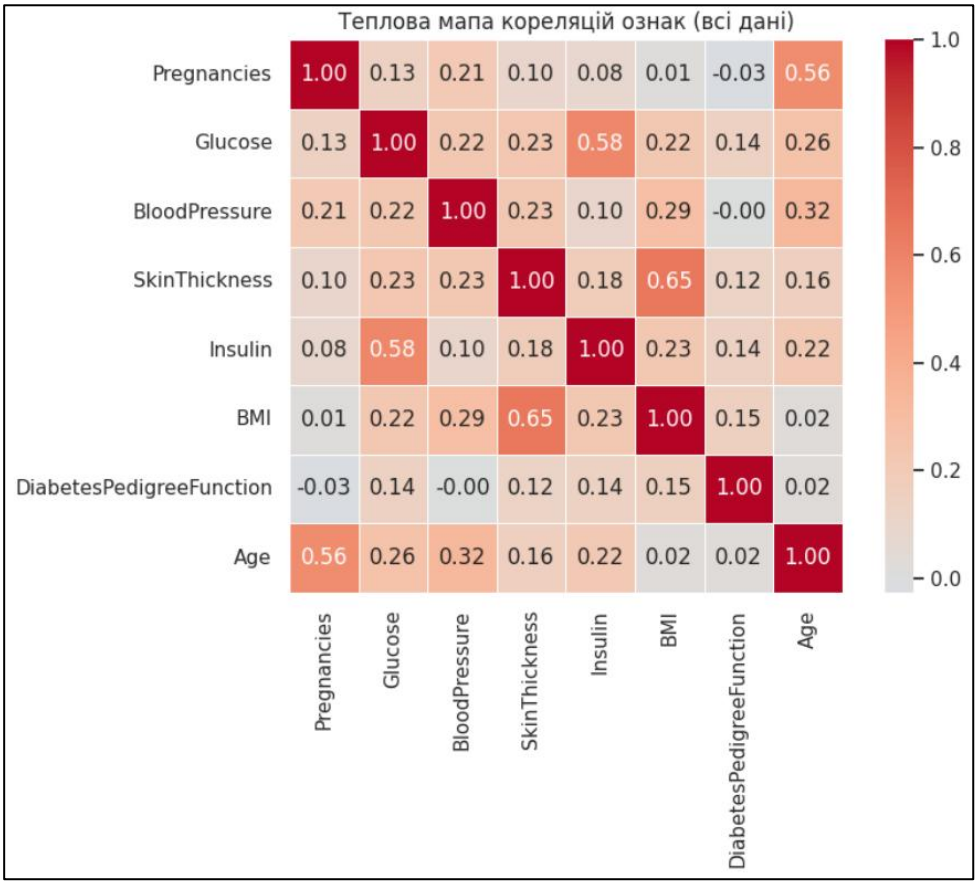

Рис. 2. Теплова мапа кореляцій ознак

На рис. 2 видно сильну кореляцію між SkinThickness і BMI (r = 0.65), між Insulin і Glucose (r = 0.58), а також між Pregnancies і Age (r = 0.56). На підставі аналізу кореляцій та медичної доцільності з моделі виключено SkinThickness, Insulin і Pregnancies. Для заповнення залишкових пропусків застосовано KNN-імпутацію окремо на тренувальній і тестовій вибірках, що запобігло витоку даних.

Фінальний набір ознак склав 5 змінних: Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, Age. Поділ на тренувальну та тестову вибірки виконано у співвідношенні 80/20: X_train — 579 записів, X_test — 145 записів.

Побудова та порівняння байєсівських моделей

Базова модель GaussianNB [3], що припускає нормальний розподіл ознак у межах кожного класу, показала на тестовій вибірці Accuracy = 0.7379, F1-score = 0.5957, ROC-AUC = 0.7440. Модель хибно класифікувала 38 із 145 зразків (26.2%), при цьому Sensitivity склала лише 0.56 при Specificity = 0.8316, що свідчить про тенденцію до пропуску позитивних випадків діабету.

Для наближення розподілів до нормального застосовано log1p-перетворення. Після перетворення BMI статистично наблизився до нормального (p = 0.346), тоді як Glucose, DiabetesPedigreeFunction та Age залишились ненормальними (рис. 3). Незважаючи на це, GaussianNB (Log1p) продемонстрував покращення: F1-score зріс до 0.6200, ROC-AUC — до 0.7669, а Sensitivity — до 0.62.

Для застосування CategoricalNB неперервні ознаки дискретизовано на бін-інтервали. Досліджено дві стратегії — рівномірну (uniform) та квантильну (quantile) — при кількості бінів від 2 до 10. Оптимальною виявилась конфігурація з 6 бінами та рівномірною стратегією (alpha = 0.1): Accuracy = 0.7310, F1-score = 0.6286, ROC-AUC = 0.7856. Квантильна стратегія при тих самих параметрах дала дещо нижчі результати: Accuracy = 0.7034, ROC-AUC = 0.7561. Перевагою CategoricalNB (Uniform) є вищий ROC-AUC порівняно з базовим GaussianNB (+0.042) та краща Sensitivity = 0.66, що є важливим у медичному контексті. Нижче можна переглянути графічно зображені матриці помилок для усіх чотирьох моделей (рис. 3).

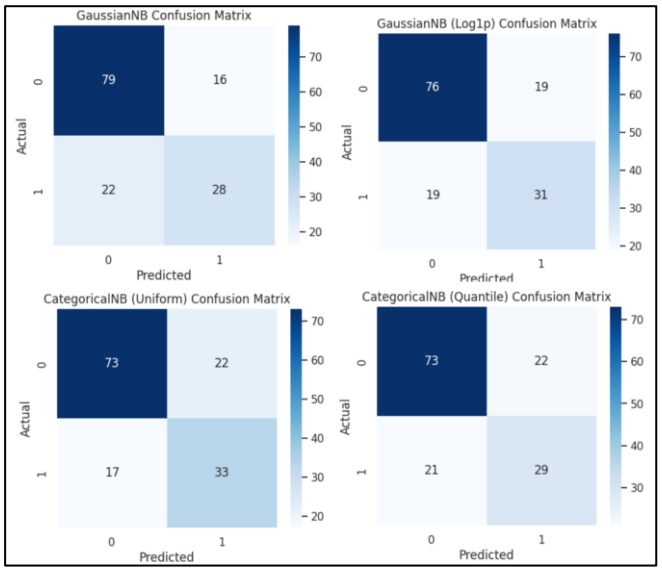

Рис. 3. Матриці помилок для GaussianNB, GaussianNB Log1p, CategoricalNB Uniform, CategoricalNB Quantile

Експерименти з покращення моделей

З метою підвищення якості прогнозування проведено серію експериментів. Підбір гіперпараметра var_smoothing для GaussianNB методом крос-валідації дав оптимальне значення 3.73×10^{-4}, що забезпечило Accuracy = 0.7586, F1-score = 0.6465, ROC-AUC = 0.7587. Досліджено також BernoulliNB після бінаризації ознак (Accuracy = 0.6690) та вплив повернення раніше виключених ознак (SkinThickness, Insulin, Pregnancies), що призвело до погіршення результатів для більшості моделей через зашумленість цих ознак.

Ансамблевий підхід Stacking на основі байєсівських базових класифікаторів дав Accuracy = 0.7586, ROC-AUC = 0.7771. Stacking із 9 різнотипних моделей (LogisticRegression, KNN, SVC, DecisionTree, AdaBoost, GradientBoosting, GaussianNB, ExtraTreesClassifier) забезпечив Accuracy = 0.7655, проте Sensitivity знизилась до 0.56. Найвищих результатів досягнуто за допомогою Voting Ensemble: Accuracy = 0.7931, ROC-AUC = 0.8055. Всі вище згадані значення метрик моделей, розроблених в рамках дослідження, можна переглянути на рис. 4.

Фінальна зведена таблиця результатів моделей

	Модель	Accuracy	F1-score	ROC-AUC	Sensitivity	Specificity
0	GaussianNB	0.7379	0.5957	0.7440	0.5600	0.8316
1	GaussianNB (Log1p)	0.7379	0.6200	0.7669	0.6200	0.8000
2	GaussianNB (var_smoothing)	0.7586	**0.6465**	0.7587	0.6400	0.8211
3	CategoricalNB (Uniform)	0.7310	0.6286	0.7856	0.6600	0.7684
4	CategoricalNB (Quantile)	0.7034	0.5743	0.7561	0.5800	0.7684
5	Stacking (Bayes ensemble)	0.7586	**0.6465**	0.7771	0.6400	0.8211
6	CategoricalNB Uni (4 ознаки)	0.6759	0.5913	0.7546	**0.6800**	0.6737
7	RandomForest [референс]	0.7241	0.6000	0.7695	0.6000	0.7895
8	Stacking (9 моделей)	0.7655	0.6222	0.7742	0.5600	0.8737
9	Voting Ensemble (9 моделей)	**0.7931**	0.6429	**0.8055**	0.5400	**0.9263**

Рис. 4. Фінальна зведена таблиця результатів усіх моделей

Висновки

У роботі побудовано та досліджено систему прогнозування ризику діабету на основі байєсівських методів класифікації. Виконано повний цикл обробки даних: виявлення та заміна медично неможливих нулів (374 записи по Insulin, 227 — по SkinThickness), відбір 5 найбільш інформативних ознак на основі кореляційного аналізу, KNN-імпутація пропущених значень.

Серед байєсівських моделей найкращий збалансований результат показав GaussianNB з підібраним var_smoothing: Accuracy = 0.7586, F1-score = 0.6465, ROC-AUC = 0.7587, Sensitivity = 0.64. CategoricalNB з рівномірною дискретизацією (6 бінів) забезпечив найвищий ROC-AUC серед байєсівських класифікаторів — 0.7856 при Sensitivity = 0.66, що є критично важливим показником для медичної діагностики, оскільки мінімізує пропуск хворих пацієнтів. Voting Ensemble із 9 різнотипних моделей досяг найвищих показників у роботі: Accuracy = 0.7931 та ROC-AUC = 0.8055, перевищивши базовий GaussianNB на 5.5 та 6.2 відсоткових пунктів відповідно.

Таким чином, байєсівські класифікатори є ефективним і обчислювально економним інструментом для медичної діагностики: час навчання GaussianNB складає менше 0.004 секунди. Для підвищення Sensitivity в клінічних застосуваннях рекомендується використовувати CategoricalNB з рівномірною дискретизацією або ансамблеві методи з порогом класифікації, скоригованим відповідно до медичних вимог.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Pima Indians Diabetes Database [Електронний ресурс]. — Режим доступу: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
2. Scikit-learn Documentation: Preprocessing [Електронний ресурс]. — Режим доступу: https://scikit-learn.org/stable/modules/preprocessing.html
3. Scikit-learn Documentation: Naive Bayes [Електронний ресурс]. — Режим доступу: https://scikit-learn.org/stable/modules/naive_bayes.html
4. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. — O'Reilly Media, 2022. — 851 p.

Янковчучк Михайло Ігорович – студент групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: mykhailoyanki@gmail.com.

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

Yankovchuk Mykhailo I. – student of Faculty of Intelligent Information Technology and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail: mykhailoyanki@gmail.com.

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com.