

БАЙЄСІВСЬКЕ МОДЕЛЮВАННЯ ДЛЯ ОЦІНЮВАННЯ РИЗИКІВ ВИНИКНЕННЯ РАКУ ЛЕГЕНІВ НА ОСНОВІ АНАЛІЗУ МЕДИЧНИХ ДАНИХ

Вінницький національний технічний університет

Анотація

У цій роботі розглядаються методи байєсівського моделювання для прогнозування ризику розвитку онкологічних захворювань. Основна увага приділяється алгоритмам Gaussian Naive Bayes (GaussianNB) та Bayesian Ridge, які були протестовані на наборі даних із Kaggle. Проведено аналіз їхньої ефективності та порівняння результатів з метою визначення найбільш придатного підходу для медичних прогнозів.

Ключові слова: Байєсівські моделі, GaussianNB, Bayesian Ridge, прогнозування раку, машинне навчання.

Abstract

This paper examines Bayesian modeling methods for predicting the risk of developing cancer. The main focus is on the Gaussian Naive Bayes (GaussianNB) and Bayesian Ridge algorithms, tested on a Kaggle dataset. An analysis of their effectiveness was conducted, and the results were compared to determine the most suitable approach for medical predictions.

Keywords: Bayesian models, GaussianNB, Bayesian Ridge, cancer prediction, machine learning.

Вступ

Байєсівські моделі широко використовуються в задачах класифікації та прогнозування завдяки їхній здатності працювати з невеликими вибірками та ефективно обробляти ймовірнісні залежності між ознаками. Однією з головних переваг є їхня інтерпретованість та можливість врахування апостеріорної ймовірності при прийнятті рішень. У медицині такі моделі особливо корисні, оскільки дозволяють оцінювати ризики захворювань та робити прогнози на основі обмежених або нерівномірно розподілених даних.

Метою цієї роботи є аналіз продуктивності алгоритмів GaussianNB і Bayesian Ridge при прогнозуванні ризику розвитку раку на основі реальних медичних даних.

Формування датасету

У цьому дослідженні використано відкритий набір даних Lung Cancer Risk & Trends Across 25 Countries, автором якого є Ankush Panday, набір має відкритий доступ для загального використання на платформі Kaggle [6]. Для побудови моделі було відібрано низку ключових змінних, що охоплюють демографічні, поведінкові та екологічні фактори, всього обрано 11 атрибутів та 15000 записів. До демографічних параметрів належать розмір популяції, вік та стать, тоді як серед змінних, пов'язаних із курінням, враховано статус курця, кількість років паління, середню кількість викурених цигарок на день, а також пасивне куріння. Окрему групу складають екологічні та професійні чинники, що включають рівень забруднення повітря, професійну експозицію до шкідливих речовин і вплив забруднення в приміщеннях (рис.1).

Цільовою змінною моделі є наявність або відсутність діагностованого раку легень, де значення 0 вказує на його відсутність, а 1 — на підтверджений діагноз. Для аналізу залежності між цими факторами ризику та ймовірністю розвитку захворювання застосовано метод Байєсівської логістичної регресії, що дозволяє оцінити вплив кожного предиктора та отримати розподіл ймовірностей для прогнозування ризику [1, 4].

	Population_Size	Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_Day	Passive_Smoker	Air_Pollution_Exposure	Occupational_Exposure	Indoor_Pollution
0	1400	80	0	1	30	29	0	0	1	0
1	84	53	0	0	0	0	1	0	1	0
2	128	47	0	1	12	6	1	1	0	0
3	273	39	1	0	0	0	0	0	0	0
4	59	44	1	0	0	0	1	1	1	0
...	...	...	...	...	...	...	...	...	...	...
9995	125	78	1	0	0	0	1	1	1	0
9996	65	80	0	0	0	0	1	1	0	1
9997	125	81	0	0	0	0	1	1	0	0
9998	225	54	0	1	3	10	0	1	1	0
9999	213	66	0	0	0	0	0	0	1	0

Рис. 1. Приклад отриманих даних про пацієнтів

Проведено попереднє очищення даних та розвідувальний аналіз даних. Побудовано матрицю кореляції змінних (рис.2).

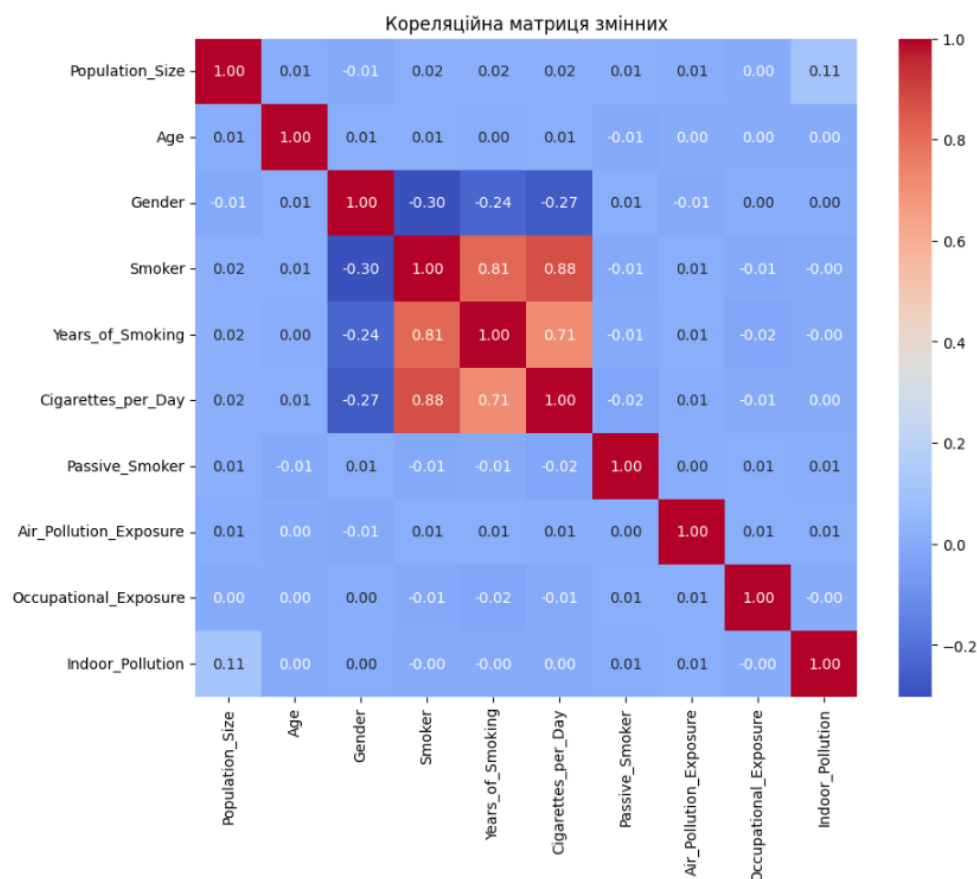


Рис. 2. Матриця кореляції

Аналіз кореляційної матриці виявив тісні взаємозв'язки між факторами, пов'язаними з палінням. Зокрема, спостерігається висока кореляція між статусом курця та кількістю років паління ($r = 0.81$), а також між статусом курця та середньою кількістю викурених сигарет на день ($r = 0.88$). Крім того, помітний зв'язок між тривалістю паління та щоденною кількістю викурених сигарет ($r = 0.71$). Такі результати є цілком очікуваними, адже зі збільшенням стажу куріння зазвичай зростає і загальна кількість спожитого тютюну [2]. Це підкреслює закономірність між інтенсивністю паління та тривалістю його впливу на організм.

Результати дослідження

У цьому дослідженні порівнюються дві Байєсівські моделі: Bayesian Ridge Regression та Gaussian Naive Bayes (GaussianNB). Перша з них є регуляризованою лінійною моделлю, яка добре підходить для прогнозування в умовах високої кореляції між змінними. Її перевага полягає у здатності ефективно працювати з шумними та варіабельними даними, а також оцінювати ймовірнісний розподіл параметрів, що робить її стійкішою до переобчислення коефіцієнтів при зміні навчальної вибірки [2, 4].

Натомість Gaussian Naive Bayes є простим Байєсівським класифікатором, який припускає нормальний розподіл ознак. Він є ефективним у ситуаціях, коли змінні незалежні, проте може втрачати точність за наявності значної кореляції між ними. Завдяки своїй простоті модель легко інтерпретується, однак на складних медичних даних, які містять залежні фактори, її ефективність може бути обмеженою.

Для навчання та тестування моделей було використано 80% даних для навчальної вибірки та 20% – для тестової. Оскільки Bayesian Ridge є лінійною моделлю, перед її застосуванням було проведено масштабування неперервних змінних, що дозволило покращити її ефективність. Для оцінки якості моделей використовувалися метрики точності, precision, recall та F1-score, що дозволило отримати об'єктивну картину їхньої продуктивності. Реалізація моделей та увесь код можна подивитися в Kaggle ноутбукі [5].

За результатами роботи моделі GaussianNB побудовано матрицю помилок (рис.3).

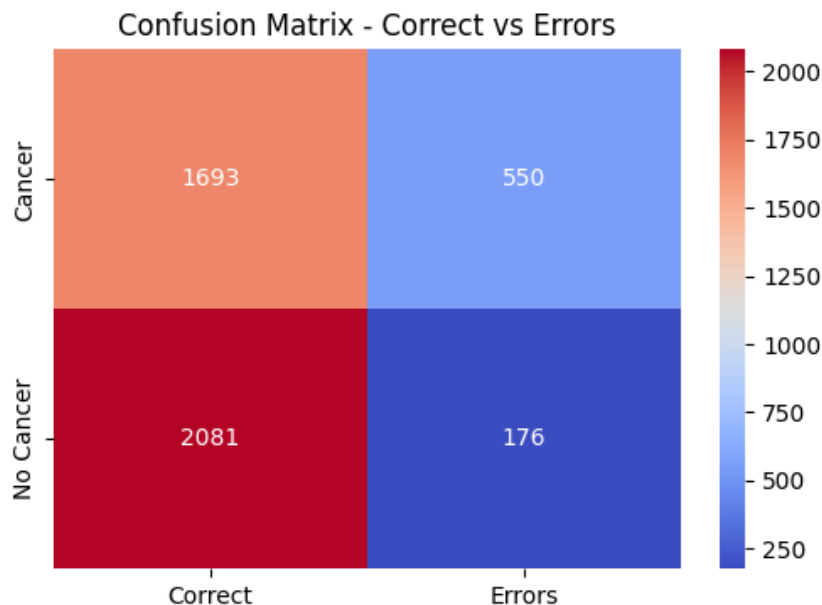


Рис. 3. Матриця помилок GaussianNB

Матриця помилок для GaussianNB (рис.3) виявила наступні характеристики:

- Правильно передбачено рак: 1693 випадків
- Помилково передбачено відсутність раку у хворих пацієнтів: 550 випадків
- Правильно передбачено відсутність раку: 2081 випадків
- Помилково передбачено наявність раку у здорових пацієнтів: 176 випадків

За результатами роботи моделі Bayesian Ridge побудовано матрицю помилок (рис.4).

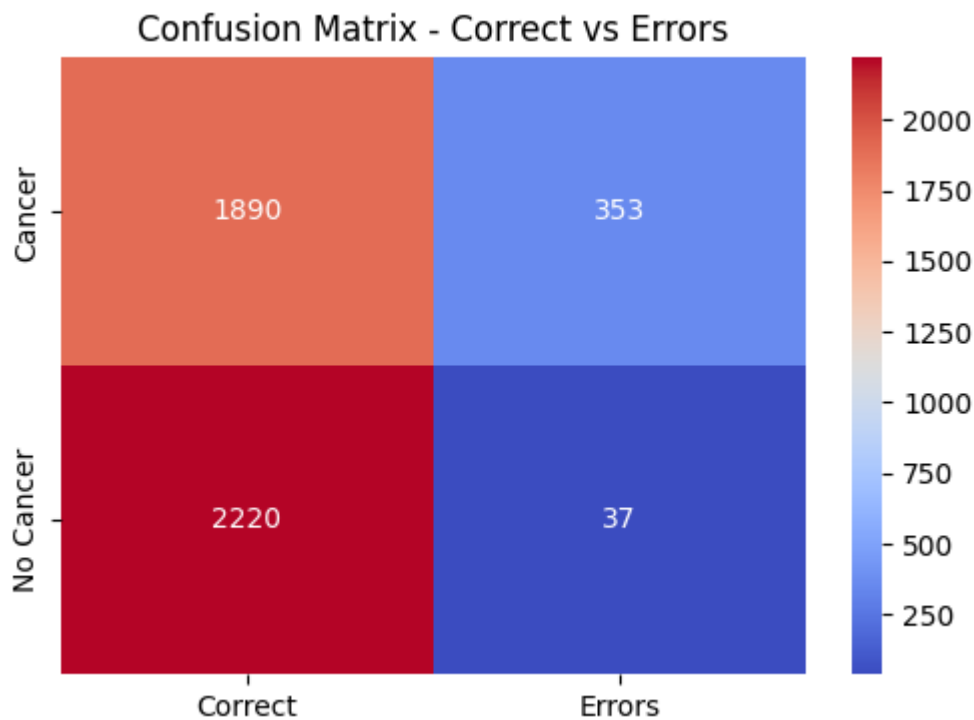


Рис. 4. Матриця помилок Bayesian Ridge

Для Bayesian Ridge (рис.4) спостерігалось покращення точності:

- Правильно передбачено рак: 1890 випадків
- Помилково передбачено відсутність раку у хворих пацієнтів: 353 випадків
- Правильно передбачено відсутність раку: 2220 випадок
- Помилково передбачено наявність раку у здорових пацієнтів: 37 випадків

Ці результати свідчать про те, що Bayesian Ridge значно перевершує GaussianNB за всіма основними метриками. Його загальна точність (91.3%) вища, ніж у GaussianNB (83.87%), що свідчить про кращу адаптацію до даних. Recall для позитивного класу («є рак») у Bayesian Ridge становить 84% проти 75% у GaussianNB, що означає меншу ймовірність пропущених випадків захворювання. Крім того, Bayesian Ridge має вищий precision для класу «є рак» (98% проти 91%), що вказує на більш точне визначення хворих пацієнтів. Загалом Bayesian Ridge демонструє кращий баланс між precision і recall, роблячи його більш придатним для медичних прогнозів, тоді як GaussianNB через високий рівень хибнонегативних результатів менш ефективний у цій сфері.

Висновки

Результати дослідження демонструють суттєву різницю в ефективності двох байєсівських підходів до прогнозування раку легень. Bayesian Ridge Regression продемонструвала високу точність (0.9133), що свідчить про її здатність ефективно працювати з медичними даними, навіть за умов значної варіабельності та кореляції між ознаками. Її класифікаційний звіт показує високу точність передбачення як для позитивного, так і для негативного класу, що особливо важливо для медичних застосувань, де критичною є правильна ідентифікація пацієнтів із ризиком розвитку хвороби.

Натомість Gaussian Naive Bayes продемонструвала нижчу загальну точність (0.8387), що, ймовірно, пов'язано з припущенням про незалежність змінних, яке в даному випадку не відповідає структурі даних. Аналіз матриці помилок вказує на вищу кількість неправильно класифікованих випадків у порівнянні з Bayesian Ridge, що може бути наслідком впливу корельованих факторів, які дана модель не враховує належним чином. Попри це, вона демонструє добрий рівень recall для класу «рак», що означає, що більшість пацієнтів із позитивним діагнозом були правильно ідентифіковані.

Отримані результати підтверджують, що Bayesian Ridge Regression є більш адаптованою до завдання прогнозування на основі комплексних медичних даних, зокрема завдяки врахуванню кореляцій між змінними та меншій чутливості до розподілу даних. Водночас GaussianNB залишається корисною моделлю, особливо у випадках, коли простота, інтерпретованість та швидкість розрахунків є визначальними критеріями. Однак для більш точної діагностики та зменшення кількості хибнопозитивних та хибнонегативних випадків Bayesian Ridge Regression виявляється значно ефективнішою.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Rodinkova V., Yuriev S., Mokin V., Kryvopustova M., Shmundiak D., Bortnyk M., Kryzhanovskiy 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. S.P. Bhargav; S. Om Prakash; S. Hariharasudhan; P. Tamilselvi. Impact of PCA on Lung Cancer Dataset Classification: A Comparative Analysis of Machine Learning Models. Published in International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024, [Електронний ресурс]. – Режим доступу: <https://ieeexplore.ieee.org/abstract/document/10533485>
3. Mohammad Shafiquzzaman Bhuiyan, Imranul Kabir Chowdhury, Mahfuz Haider, Afjal Hossain Jisan, Rasel Mahmud Jewel, Rumana Shahid, Mst Zannatun Ferdus, & Siddiqua, C. U. (2024). Advancements in Early Detection of Lung Cancer in Public Health: A Comprehensive Study Utilizing Machine Learning Algorithms and Predictive Models. Journal of Computer Science and Technology Studies, 6(1), 113-121, [Електронний ресурс]. – Режим доступу: <https://doi.org/10.32996/jcsts.2024.6.1.12>
4. Zhong, L., Yang, F., Sun, S. et al. Predicting lung cancer survival prognosis based on the conditional survival bayesian network. BMC Med Res Methodol 24, 16 (2024), [Електронний ресурс]. – Режим доступу: <https://doi.org/10.1186/s12874-023-02043-y>
5. Roman Bralatan notebook [Електронний ресурс] – Режим доступу: <https://www.kaggle.com/code/romantick/notebook4acec5e3be>
6. Lung Cancer Risk & Trends Across 25 Countries Dataset. Kaggle. 2023 [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/ankushpanday1/lung-cancer-risk-and-trends-across-25-countries>

Бралатан Роман Андрійович – аспірант 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. - postgraduate 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