УДК 004.021
DOI https://doi.org/10.32782/2521-6643-2023.2-66.3

Бойко Н. І., кандидат економічних наук, доцент,
доцент кафедри систем штучного інтелекту
Національного університету «Львівська політехніка»
ORCID: 0000-0002-6962-9363

Курило В., студент 4-го курсу кафедри систем штучного інтелекту
Національного університету «Львівська політехніка»

АЛГОРИТМ КЛАСИФІКАЦІЇ МЕДИЧНИХ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ ОНКОЛОГІЇ

У даній роботі наведено результати досліджень застосування Логістичної регресії та Дерева рішень з використання алгоритму PCA в задачі виявлення та прогнозуванні онкології. Було проаналізовано проблему та актуальність даного дослідження. Проаналізовано різноманітні літературні джерела та методи машинного навчання. Проведений детальний аналіз обраних методів, а також розглянути їх математичні моделі. Було проведено тренування відповідних моделей та ряд експериментів для обрання найкращих параметрів на обраних двох наборах даних, які були детально проаналізовані в даній роботі. Наведено результати точності моделей та побудовані відповідні метрики, такі як Classification report, Confusion Matrix, Roc-curve. Також були проведені експерименти для покращення точності моделей з використання алгоритму PCA. В результаті були отримані набагато кращі результати у випадку з другим набором даних, але з першим покращити точність не вдалося. Після проведення експериментальної частини було детально проаналізовано отримані результати та наведені відповідні гістограми для кожного з наборів даних з отриманими результатами. Дане дослідження доводить, що алгоритм PCA краще використовувати, тоді коли наявний набір даних з великою кількістю ознак. В результаті дослідження були отримані хороші результати у задачі виявлення та прогнозування онкології та наведено цінність даного дослідження з описаними висновками. В роботі проводиться оцінка результатів за допомогою різних метрик, таких як точність та чутливість, і результати порівнюються з іншими методами аналізу та класифікації. Доведено, що ці методи можуть вдосконалити процес діагностики онкології, сприяти зменшенню помилкових класифікацій та сприяти ранньому виявленню хвороби.

Ключові слова: *алгоритм, машинне навчання, онкологія, виявлення, прогнозування.*

Boyko N. I., Kurylo V. Medical data classification algorithm for oncology prediction

The article presents the results of research on the application of Logistic Regression and Decision Tree with the use of PCA algorithm in the task of cancer detection and prediction. The problem and relevance of this research are analyzed. Various literature sources and machine learning methods are reviewed. A detailed analysis of the chosen methods is conducted, along with their mathematical models. Training of respective models and a series of experiments are carried out to select the best parameters on two selected datasets, which are thoroughly analyzed in the study. The accuracy of the models is evaluated, and corresponding metrics such as Classification report, Confusion Matrix, and Roc-curve are constructed. Additionally, experiments are conducted to enhance the accuracy of the models using the PCA algorithm. As a result, significantly improved outcomes are achieved with the second dataset, while the accuracy improvement is not achieved with the first dataset. After the experimental phase, the obtained results are analyzed in detail, and corresponding histograms with the results are provided. This research demonstrates that the PCA algorithm is better utilized when dealing with datasets with a large number of features. Overall, the study yields promising results in cancer detection and prediction, and the value of this research is highlighted with the described conclusions. The paper evaluates the results using various metrics, such as accuracy and sensitivity, and compares the results with other analysis and classification methods. It has been proven that these methods can improve the process of oncology diagnosis, contribute to the reduction of false classifications and contribute to the early detection of the disease.

Key words: *algorithm, machine learning, oncology, detection, prediction.*

Постановка проблеми. Машинне навчання активно розвивається та уже давно набуло неабиякої важливості у сучасному світі, в тому числі в медицині. В наші дні важко уявити лікарню без використання сучасних технологій AI. За допомогою яких встановлюють діагнози, прогнозують розвиток хвороб, аналізують стан пацієнта і багато іншого. В основі роботи МН в медицині лежить аналіз великих медичних даних, завдяки обробці яких та з використанням різноманітних методів МН ми можемо покращити роботу медицини, оскільки алгоритми МН можуть виявити ті відхилення, які не може виявити людини.

Темою роботи є прогнозування онкології з використання методів машинного навчання, оскільки цей діагноз майже не виліковний та мало хто його виявляє на ранніх стадіях, через що помирає велика кількість людей. Цей напрямок був обраний, тому що рак є основною проблемою охорони здоров'я та провідною глобальною причиною смерті, де скринінг, діагностика, прогнозування, оцінка виживаності та лікування раку та заходи контролю все ще є серйозною проблемою.

Звідси видно, що прогнозування онкології з використанням методів машинного навчання є важливим та актуальним напрямком досліджень в сучасному світі. Онкологічні захворювання становлять серйозну загрозу для здоров'я та життя людей, а раннє виявлення та прогнозування можуть значно покращити їх виживання та результати лікування. Застосування методів машинного навчання дозволяє аналізувати великі обсяги медичних даних та ідентифікувати фактори ризику та ознаки онкології, які людина може пропустити. Це відкриває нові можливості для покращення стратегій профілактики, діагностики та лікування раку. Прогнозування онкології з використанням методів машинного навчання може використовуватися як інструмент для раннього виявлення та індивідуалізації лікування, що сприяє покращенню результатів та збільшенню шансів на одужання у пацієнтів з онкологічними захворюваннями.

Для подальшого аналізу було обрано два методи, а саме логістичну регресію та дерево рішень, а також алгоритм PCA для зменшення розмірності даних та покращення точності моделі. Дані методи були обрані, оскільки вони є зрозумілі і прості у реалізації, а також вони мають свої унікальні переваги та можуть принести цінну інформацію для прогнозування онкології.

Аналіз останніх досліджень і публікацій. Актуальність дослідження полягає в тому, що прогнозування онкології є важливою задачею в медицині, а використання логістичної регресії та дерева рішень дозволяє отримати цінну інформацію для класифікації та прогнозування онкологічних захворювань. Додатково, використання алгоритму PCA для зменшення розмірності даних може покращити точність моделі, знизити шум та покращити роботу зі збалансованими ознаками.

В роботі [5] розглядаються обрані методи разом з алгоритмом PCA мають потенціал покращити точність та надійність моделі прогнозування онкології, що робить дане дослідження актуальним та новизною у використанні цих методів для проблеми онкології.

У статті [1] автори проводять докладний огляд робіт, які використовують машинне навчання для класифікації та прогнозування результатів пацієнтів з використанням електронних медичних записів. Вони аналізують різні алгоритми глибокого навчання, набори даних, використані показники та результати досліджень. Тут показані, такі речі як: Лінійна регресія, градієнтний спуск, SVM, дерево рішень та інші. Головна перевага статті [1], що це все є пояснення, як на простих прикладах, так і у раковій діагностиці, що значно покращує розуміння цих методів в подальших дослідженнях. Стаття [1] є корисною у дослідженні, оскільки надає вичерпний огляд методів та моделей машинного навчання, які застосовуються для прогнозування результатів пацієнтів на основі електронних медичних записів. Вона допомагає зрозуміти різні підходи, що застосовуються в цій області, а також виявити потенційні переваги та виклики, пов'язані з використанням машинного навчання на основі електронних медичних записів.

У роботі [2] автори досліджують різні алгоритми та моделі машинного навчання, що застосовуються для прогнозування раку. Вони використовують набори даних, що містять клінічні та генетичні характеристики пацієнтів, для навчання та тестування моделей. У статті [2] автори розглядають різні методи машинного навчання, такі як логістична регресія, дерева рішень, метод опорних векторів (SVM) та наївний баєсів класифікатор. Вони порівнюють ефективність цих методів у прогнозуванні ракових захворювань та визначають найбільш точні та надійні підходи. Дослідження показує, що використання машинного навчання може допомогти вдосконалити прогнозування раку та сприяти ранній діагностиці цього захворювання. Результати цієї роботи є корисні для подальшого розвитку методів прогнозування та діагностики раку на основі машинного навчання.

Стаття [3] є оглядом використання методів машинного навчання в передбаченні раку. У статті автори розглядають різні алгоритми та моделі машинного навчання, що використовуються для аналізу клінічних та генетичних даних у ракових дослідженнях. Стаття описує різні методи машинного навчання, такі як логістична регресія, метод опорних векторів (SVM), дерева рішень, нейронні мережі та ансамблеві моделі, і вказує на їхню ефективність у прогнозуванні результатів ракових захворювань. Ця стаття є корисною для дослідження, оскільки надає зрозумілий огляд різних методів машинного навчання, що застосовуються у прогнозуванні раку. Вона надає базові знання про різні алгоритми та їхні можливості, що допоможе у виборі підходів для дослідження. Крім того, стаття також вказує на потенційні виклики та перспективи використання машинного навчання в ракових дослідженнях.

Стаття [4] фокусується на оцінці та порівнянні продуктивності алгоритмів зменшення розмірності в моделях машинного навчання для прогнозування раку. Автори проводять комплексний аналіз різних методів зменшення розмірності, які використовуються в поєднанні з алгоритмами машинного навчання для прогнозування раку. У статті [4] досліджуються різні методи зменшення розмірності, основний з них – аналіз головних компонентів (PCA). Автори оцінюють ефективність цих алгоритмів у зменшенні розмірності наборів даних, пов'язаних із раком, і покращенні продуктивності моделей машинного навчання в прогнозуванні раку. Ця

стаття є цінною для дослідження, оскільки вона дає уявлення про ефективність і придатність різних алгоритмів зменшення розмірності для прогнозування раку. Це допомагає зрозуміти вплив методів зменшення розмірності на точність і ефективність моделей машинного навчання в дослідженнях, пов'язаних із раком.

Метою цього дослідження є оцінка та порівняння методів машинного навчання на прогнозування онкології. Це дослідження ставить за мету визначити оптимальні методи для класифікації та прогнозування ракових захворювань на основі доступних клінічних та генетичних даних. Результати дослідження можуть сприяти покращенню точності та надійності моделей прогнозування онкології та клінічних стратегій у боротьбі з цими хворобами.

Виклад основного матеріалу. Математична постановка задачі полягає у побудові моделей за допомогою логістичної регресії, дерева рішень та алгоритму PCA для прогнозування онкологічних захворювань. Нехай маємо набір незалежних змінних $X = \{X_1, X_2, \ldots, X_p\}$, де p – кількість ознак, і залежну змінну Y, що вказує на наявність загрози або її відсутність онкологічного захворювання. Метою є побудова моделі, яка здатна класифікувати нові спостереження в залежності від їхніх ознак та зробити прогноз наявності онкологічного захворювання. Для даної задачі прогнозування онкології за допомогою логістичної регресії, математична модель може бути описана наступним чином:

1. Припустимо, що маємо набір клінічних та генетичних ознак для N прикладів, які позначимо як $x_i \ (1 \le i \le N)$. Кожен x_i є вектором ознак розмірності d.

2. Для кожного прикладу x_i, маємо відповідну мітку класу y_i, де y_i=1, якщо приклад відноситься до позитивного класу (наявність загрози онкологічного захворювання), і y_i = 0, якщо приклад належить до негативного класу (відсутність загрози онкологічного захворювання).

3. Логістична регресія моделює ймовірність P($y_i = 1 \mid x_i$) залежно від ознак x_i за допомогою логістичної функції – сигмоїди (Формула 1):

$$P(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-z_i)}, \tag{1}$$

де $z_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \ldots + \beta\_d * x_{id}$

$\beta_0, \beta_1, \ldots, \beta\_d$ є параметрами моделі, які потрібно знайти.

4. Задача полягає у знаходженні оптимальних значень параметрів β, які мінімізують функцію втрат та максимізують ймовірність правильного класифікації. Часто використовується метод максимальної правдоподібності або метод градієнтного спуску для знаходження цих оптимальних значень.

5. Після знаходження оптимальних значень параметрів β, можна використовувати модель для прогнозування класів для нових прикладів. Для прикладу якщо P(y = 1 | x) > 0.5, класифікуємо приклад як позитивний клас в іншому випадку класифікуємо його як негативний клас.

Математично, логістична регресія використовує логістичну функцію для моделювання ймовірностей класів залежно від набору ознак. Задача полягає у знаходженні оптимальних параметрів, які дозволяють максимізувати точність класифікації та прогнозування онкологічних захворювань на основі даних ознак.

Якщо говорити про математичну модель дерева рішень, то також припустимо, що маємо набір клінічних та генетичних ознак для N прикладів, які позначимо як x_i (1 <= i <= N). Кожен x_i є вектором ознак розмірності d.

1. Дерево рішень розбиває набір даних на декілька груп (вузлів) засновану на значеннях ознак. Кожен вузол має свою умову розділення, яка базується на значенні однієї з ознак. Наприклад, якщо ознака x_i > t, перейти до лівого піддерева, в іншому випадку – до правого піддерева.

2. Задача полягає у побудові оптимального дерева рішень, яке мінімізує помилку класифікації та максимізує точність прогнозування. Для цього можна використовувати різні критерії, такі як ентропію (Формула 2) або критерій Джині (Формула 3), для вибору найкращої ознаки та значення розділення в кожному вузлі.

$$E(x) = \sum_{i=1}^{n} - p_i * log_2(p_i), \tag{2}$$

де в даній формулі x – поточний стан, p – ймовірність події i стану x

$$Gini(t) = 1 - \sum_{i=1}^{n} (p(i \mid t))^2, \tag{3}$$

де t – вузол дерева рішень, $p(i|t)$ – ймовірність, що об'єкт належить класу і в умовах вузла t.

3. Дерево будується рекурсивно, розбиваючи набір даних на підмножини в кожному вузлі досягнення критерію зупинки, наприклад, максимальної глибини дерева або мінімальної кількості прикладів в вузлі.

4. Після побудови дерева, можна використовувати його для класифікації нових прикладів, пройшовши по шляху від кореня дерева до листя, де кожен лист представляє конкретний клас (позитивний або негативний).

Математично, дерево рішень розбиває набір даних на підмножини на основі значень ознак і рекурсивно будує гілки рішень досягнення критерію зупинки. Задача полягає у знаходженні оптимального дерева,

яке мінімізує помилку класифікації та максимізує точність прогнозування онкологічних захворювань на основі даних ознак.

Також наведемо математичну модель алгоритму PCA. Для даної задачі прогнозування онкології, математичний опис алгоритму PCA може бути наступним:

1. Припустимо, що маємо набір даних, який складається з N прикладів з d ознаками, які утворюють матрицю X розмірності N x d.

2. Спочатку проводиться центрування даних шляхом віднімання середнього значення кожної ознаки від відповідних значень усіх прикладів. Це забезпечує нульове середнє значення для кожної ознаки.

3. На наступному кроці обчислюється коваріаційна матриця C розмірністю d x d шляхом множення матриці X на її транспоновану версію і поділення на (N – 1). Коваріаційна матриця відображає ступінь взаємозв'язку між ознаками.

4. Виконується розклад коваріаційної матриці C за допомогою методу сингулярного розкладу (Singular Value Decomposition, SVD). Цей розклад дає нам власні значення та власні вектори коваріаційної матриці.

5. Головні компоненти обираються в порядку спадання їх власних значень, що відображає важливість кожної компоненти в поясненні дисперсії даних. Можна вибрати перші k головних компонент, які пояснюють більшу частину загальної дисперсії (наприклад, 90%).

6. Отримані головні компоненти утворюють нову матрицю зменшеної розмірності F розмірністю N x k, де кожний стовпчик представляє одну головну компоненту.

Алгоритм PCA дозволяє зменшити розмірність даних, зберігаючи при цьому більшу частину їх варіації. В даній задачі прогнозування онкології, алгоритм PCA може допомогти зменшити кількість ознак та покращити точність моделі, забезпечуючи важливу інформацію про взаємозв'язки між ознаками у зменшеному просторі.

Для подальшого аналізу було обрано два набори даних, які були завантажені з сайту Kaggle. Перший набір даних містить результати діагнозів 768 людей, де за восьми медичними ознаками прогнозуватимемо чи загрожує онкологія людині чи ні (0 або 1). Продемонструємо перших п'ять полів на Рис. 1.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Cancer_Markers	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Рис. 1. Демонстрація першого набору даних

Другий набір даних містить 1000 рядків та 26 медичних ознак за якими прогнозуватимемо три варіанти відповідей «Мала», «Середня», «Висока» загроза захворюваності на онкологію. На Рис. 2 продемонструємо перших п'ять полів.

Index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	1	2	3	4	Low
1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	2	1	7	2	Medium
2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	4	6	7	2	High
3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	5	6	7	5	High
4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	2	4	2	3	High

5 rows × 26 columns

Рис. 2. Демонстрація другого набору даних

Для подальшого використання обрані набори даних були перевірені на відсутні дані, викиди, аномальні значення та дублікати, а також були нормалізовані для покращення ефективності моделей та розділені на навчальні, валідаційні, тестувальні дані.

Спочатку, натренуємо моделі Логістичної регресії та Дерева рішень на першому наборі даних, який складається з 8 ознак. Для цього використаємо бібліотеку keras та її відповідні методи.

```
model = LogisticRegression()
```

Навчимо модель на навчальних даних та візуалізуємо наші результати на тестовій вибірці.

```
Accuracy:  0.8177083333333334
```

Точність нашої моделі досягає майже 82%, що свідчить про непогані результати, але які слід покращити. Для спочатку слід перевірити чи не відбувається перенавчання – коли на навчальних даних результати набагато вищі, ніж на тестових. Точність моделі на навчальних даних

```
Accuracy on train data:  0.7690972222222222
```

Бачимо, що точність на навчальних даних сягає 77%, що приблизно на 5% менше ніж на тестових даних і це означає те, що перенавчання не відбувається. Ще також відобразиться у звіті про класифікацію (Classification report), який містить дані про точність (precision) – відсоток правильних позитивних прогнозів відносно загальної кількості позитивних прогнозів. Відкликання (Recall) – відсоток правильних позитивних прогнозів відносно загальної кількості фактичних позитивних результатів. Оцінка F1(f1-score) – зважене гармонічне середнє значення точності прогнозів (Рис. 3).

	precision	recall	f1-score	support
0	0.82	0.93	0.87	125
1	0.82	0.61	0.70	67
accuracy			0.82	192
macro avg	0.82	0.77	0.78	192
weighted avg	0.82	0.82	0.81	192

Рис. 3. Classification report

Для кращого розуміння отриманих результатів побудуємо “Confusion Matrix”, яка показує кількість правильних та неправильних прогнозів. Зобразимо матрицю на Рис. 4.

Рис. 4. Confusion matrix

З Рис. 3 можна побачити, що модель робить 9 невірних прогнозів та 116 вірних у випадку, коли мітка повинна бути False або 0 та 26 невірних, а 41 вірну, коли мітка повинна бути True або 0.

Спробуємо покращити отримані результати використовуючи алгоритм PCA. Для цього були проведені експерименти з використанням різної кількості компонент. В результаті використовуючи PCA при різній кількості компонентів, результати були рівними при 2–5 компонентах та їх точність сягала

```
Accuracy on test data:  0.765625
Accuracy on train data:  0.7118055555555556
```

Тільки з зведенням датасету до семи компонентів точність трішки збільшилася, але не перевершила оригінальний датасет. Також зобразимо точність та криву-ROC при 7 компонентах

```
Accuracy on test data:  0.8020833333333334
Accuracy on train data:  0.7673611111111112
```

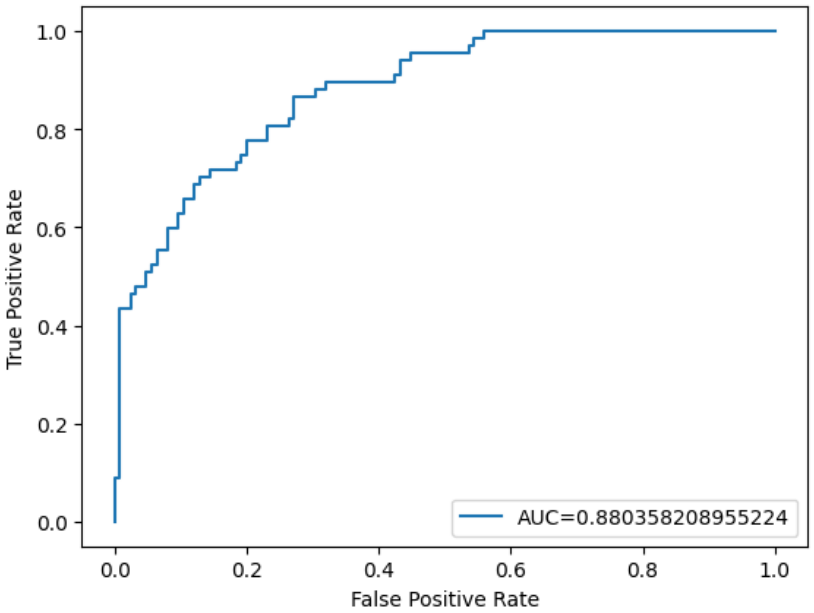

Рис. 5. ROC-Curve з даними PCA

Зробивши експерименти з використанням PCA, можна сказати, що в даному випаду він не дав очікуваних результатів, оскільки не покращив точності, а може бути корисним, тільки у візуалізації та аналізі датасету.

Наступним кроком є тренування моделі Дерева рішень. Для цього були зроблені відповідні експерименти, які показали, що модель відпрацьовує краще з критерієм Entropy та дає найкращі результати з максимальною довжиною 6.

model = DecisionTreeClassifier(criterion = 'entropy', max_length=6)

Найкраща точність 77% на тестовій вибірці досягається, коли глибина дерева рівна = 6. Також давайте подивимося на точність на тренувальній вибірці.

```
Accuracy on test data:  0.7708333333333334
Accuracy on train data:  0.8454861111111112
```

Проаналізувавши отримані результати, бачимо, що модель Логістичної регресії дає кращі результати. Але спробуємо покращити результати Дерева рішень використовуючи алгоритм PCA. Зробивши експерименти при різній кількості компонент найкраща точність була досягнута при 4 ознаках.

```
Accuracy on test data:  0.78125
Accuracy on train data:  0.8194444444444444
```

Побудуємо Classification report для даної моделі (Рис. 6).

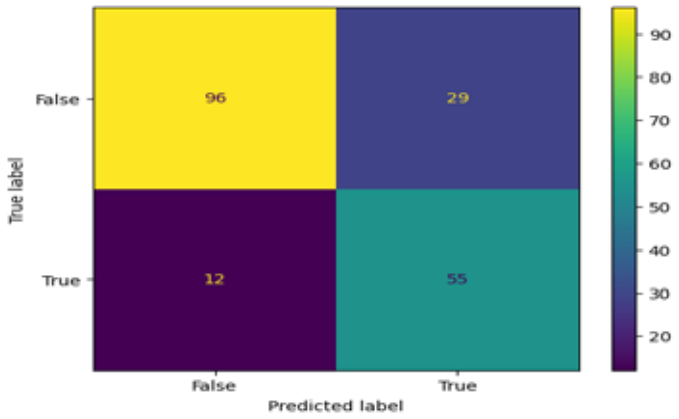

Рис. 6. Classification report

Точність на тестовій вибірці досягла 78%, що трохи покращило результати. Бачимо, що модель робить 29 помилок при класі False та 12 при класі True, що значно відрізняється від результатів логістичної регресії,

оскільки вона давала більше правильних відповідей з класом False. Але цей результат не покращив результати Логістичної регресії.

Тепер проробимо ці всі ж кроки для другого набору даних, який складається з 26 ознак. Побудуємо мультиноміальну логістичну регресію за допомогою пакета keras.

model = LogisticRegression(multi_class='multinomial')

Навчимо модель та зобразимо отриману точність на тестових та навчальних даних

```
Accuracy on test data:  0.88
Accuracy on train data:  1.0
```

Точність на тестових даних – 88%, що є доволі непогано. А на тренувальних 100%, що свідчить про можливий процес перенавчання. Це було б можливо виправити додавши більшу кількість даних, оскільки використовується, тільки одну десяту з запропонованого датасету. Але потрібно спробувати покращити ситуацію використовуючи PCA. Але спочатку ще відобразимо Confusion Matrix.

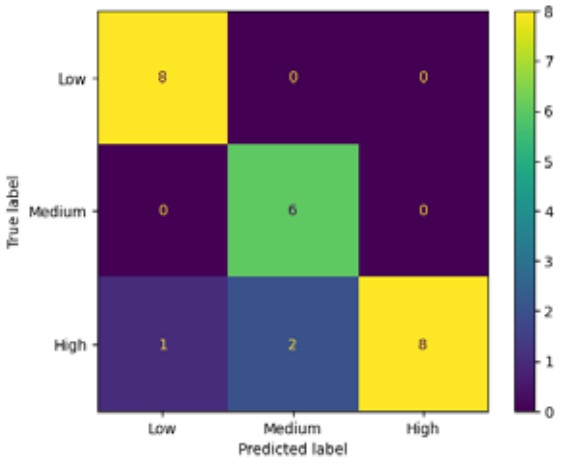

Рис. 7. Confusion Matrix

На рис. 6 показано, що модель жодного разу не помилилася на даних з вихідним значенням “Low” та “Medium”, але зробила три помилки з “High”.

Перевіримо, як модель буде себе вести з використанням PCA з різною кількістю компонент, для цього пройдемося циклом по всіх можливих варіантах та виведемо найкращий результат.

```
Accuracy on test data:  0.96
Accuracy on train data:  0.9466666666666667
```

Як бачимо, що результати значно покращились. Дана точність була досягнути з використання 3–6 компонент. Точність на тестовій вибірці значно підросла, а на тренувальній знизився. Також зобразимо результати по кожній з вихідних ознак з Confusion Matrix.

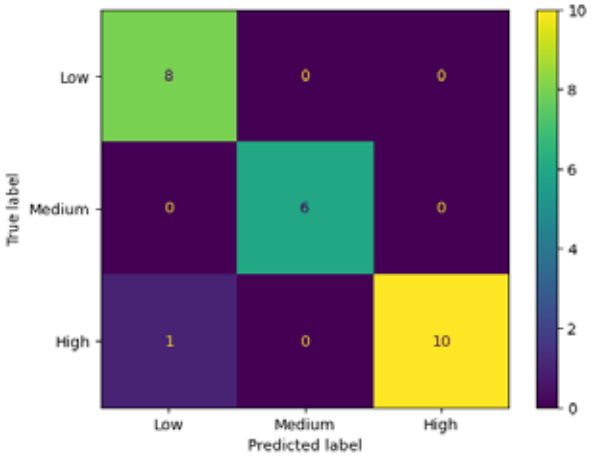

Рис. 8. Classification report

З отриманих результатів бачимо, що тепер модель працює практично ідеально і робить тільки одну помилку на тестових даних. В даному випадку PCA допоміг отримати кращі результати.

Також використаємо ще Дерево рішень і подивимося чи воно покращить нам результати для даного датасету. Для навчання та тестування використаймо цей же оброблений датасет. Як і попередній побудові дерева рішень знайдемо найоптимальнішу глибину та кількість компонент. Найкращі результати з максимальною глибиною 7. З критерієм Gini точність рівна 96%, а з Entropy точність рівна 92% без використання PCA. Використанням PCA точність значно зростає та досягає 100% в деяких випадках. Давайте побудуємо одне дерево рішень, де точність рівна 100%, а саме де кількість компонент рівна п’яти з критерієм Gini.

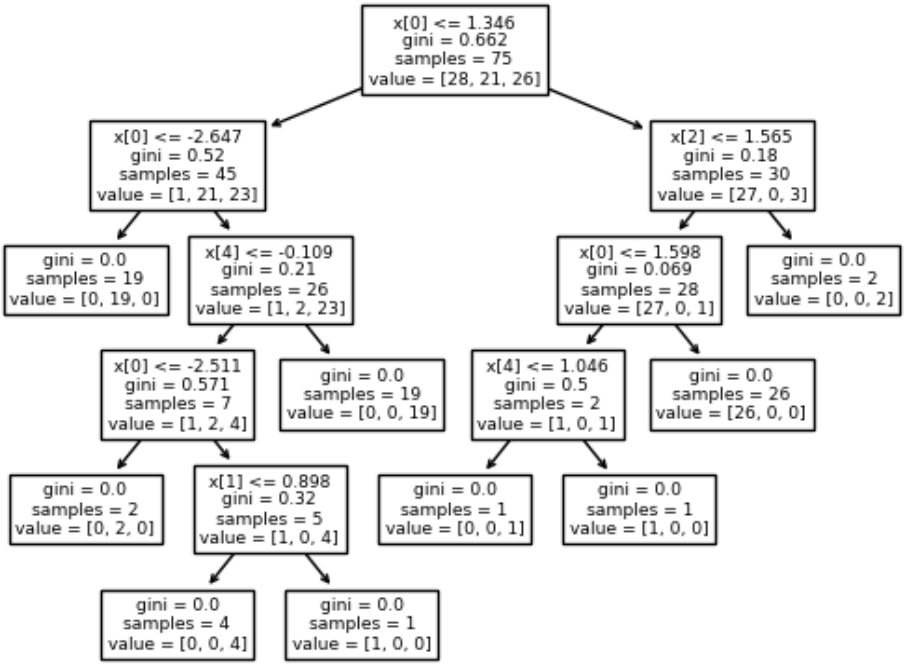

Рис. 9. Побудоване Дерево рішень

Зробивши експерименти для цього датасету, отримали хороші результати та навіть в деяких випадках дійшли точності, яка рівна 100%. В цьому нам допоміг алгоритм PCA, який став дуже корисним у випадку даного датасету.

В результаті досліджень для першого датасету вдалося досягти максимальних точностей, які зображені на Рис. 10.

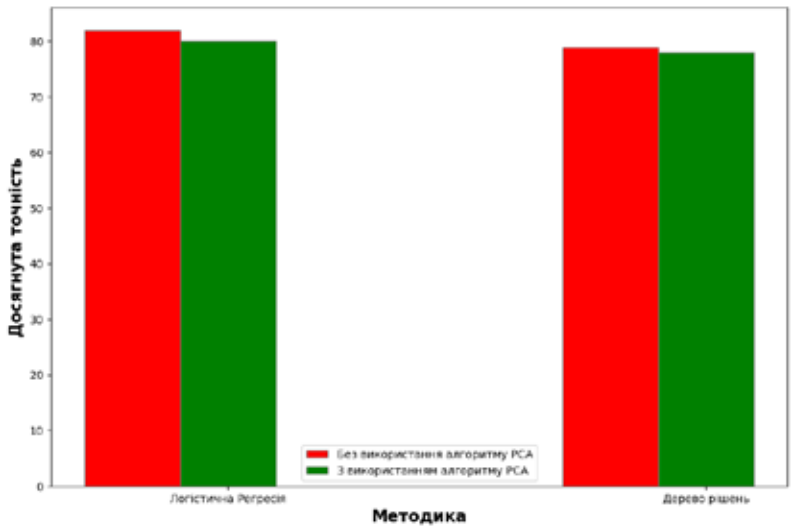

Рис. 10. Стовпчаста діаграма Досягнутої точності для першого датасету

З даної діаграми, бачимо що максимальна точність 82% з використанням біноміальної логістичної регресії. А от з використанням дерева рішень максимальна досягнута точність на тестувальних даних рівна 78%. Також покращувались результати з використання алгоритму PCA з різною кількість компонент, але в даному випадку не вдалося покращити точність моделі. Бачимо, що для цього так званого двійкового датасету найкращі результати ми отримали використовуючи біноміальну логістичну регресію, оскільки вона безпосередньо моделює ймовірність належності до одного з класів(рак загрожує або ні), а не моделює саму змінну відповіді. Даний метод є дуже корисним у використанні з двійковими даними(коли на виході є два класи), що ми довели на практиці, а також може бути особливо корисним, коли існує нелінійний зв'язок між прогнозами та ймовірністю успіху.

У випадку з другим датасетом вдалося досягти дуже хороших результатів, які візуалізуємо на Рис. 11.

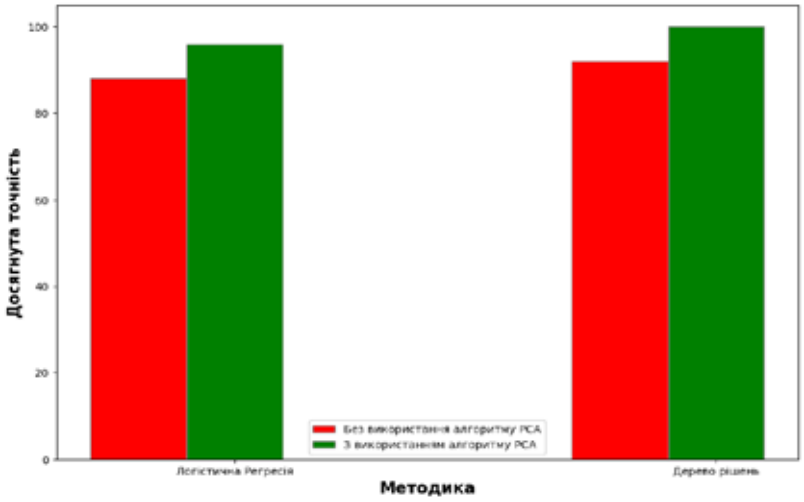

Рис. 11. Стовпчаста діаграма досягнутої точності для другого датасету

Бачимо, що у цьому випадку максимальна точність, яку вдалося досягти рівна 100% на тестовій вибірці. Такий чудовий результат ми отримали завдяки використанню методу дерева рішень з глибиною 7 та використання алгоритму PCA, а от без цього алгоритму точність моделі рівна 92%. В цьому випадку алгоритм PCA значно покращив точність моделі, так як і з використанням методу логістичної регресії, так і дерев рішень. Також була використана мультимедійна логістична регресія, точність якої сягає 88% без використання PCA та 96% з використанням. В результаті були отримані дуже хороші результати з використанням обох методів, але кращих результатів вдалося досягти з використанням методу дерев рішень, оскільки він може фіксувати нелінійні зв'язки між змінними предиктора та змінною відповіді, тоді як логістична регресія передбачає лінійний зв'язок між предикторами та логарифмом шансів змінної відповіді, а також метод дерева рішень є непараметричним методом, тобто він не вимагає припущень щодо розподілу даних або функціональної форми зв'язку між предикторами та змінною відповіді.

Якщо говорити про ефективність алгоритму PCA, то можна сказати що даний метод покращує точність, тоді коли у нас є дані великої розмірності, як це у нас є у випадку другого датасету. А от у випадку першого датасету даний алгоритм не покращує точність моделі, оскільки цей датасет не складається з великої кількості ознак. Під час роботи з даними великої розмірності кількість ознак може бути набагато більшою, ніж кількість спостережень, що може призвести до переобладнання та поганого узагальнення. Зменшуючи розмірність даних за допомогою PCA, модель може зосередитися на найважливіших функціях, що призводить до кращої продуктивності. А також даний алгоритм є корисний у випадку: 1) візуалізації даних, оскільки він може зменшити багатовимірні дані до низьковимірного простору, який можна легко візуалізувати; 2) коли у нас є корельовані функції, тобто коли функції в даних сильно корельовані одна з одною, модель може страждати від мультиколінеарності, що може призвести до нестабільних і неточних оцінок параметрів моделі. PCA може допомогти декорелювати функції та зменшити вплив мультиколінеарності, покращуючи точність моделі; 3) у зменшені шуму – зменшуючи розмірність даних за допомогою PCA, шум можна видалити, що призводить до чистішого та точнішого представлення даних; 4) прискорення обчислень: іноді зменшення розмірності даних за допомогою PCA може призвести до пришвидшення обчислень, оскільки це зменшує кількість функцій, які потрібно обробити.

Якщо порівнювати дані методи, то важко визначитися, який працює краще, оскільки, як і логістична регресія, так і дерево рішень показали хороші результати. А от який метод вибрати треба дивитися по ситуації, так як бачимо, що у нашому випадку в одному випадку себе краще проявила логістична регресія, а в другому дерева рішень. Так само у випадку використання алгоритму PCA, який у для другого датасету

підвищив точність моделі, а от для першого наоборот трішки знизив. Тому вибір потрібно робити спираючись на обраний датасет. Якщо датасет складається тільки з двох вихідних ознак, то кращі результати може показати біноміальна логістична регресія, а якщо датасет складається з великої кількості вхідних ознак то було б правильно застосовувати алгоритм PCA та використати, як і метод дерев рішень, так і логістично регресію. В загальному обидва методи можуть дати хороші результати.

Висновки та перспективи подальших досліджень. В даній роботі було проведено дослідження щодо виявлення та прогнозування раку з використанням методів машинного навчання. Проаналізувавши різноманітні літературні джерела та провівши значну кількість експериментів були отримані хороші результати. Для прогнозування онкологій були задіяні два методи машиного навчання – Логістична регресія та Дерево рішень. Завдяки яким була досягнута висока точність, яка сягає з використанням Логістичної регресії 82% та Дерева рішень 78% для першого експерименту та відповідно 96% та 100% для другого. Також був використаний алгоритм PCA, який чудово себе проявив у випадку другого експерименту, де використовувався великий датасет, оскільки він значно підвищив точності моделей приблизно на 8%, а от у випадку першого експерименту, де використовується датасет з малою кількістю ознак покращення точності не відбулося. Завдяки аналізу різних досліджень можна зробити декілька ключових висновків.

По-перше, як алгоритми логістичної регресії, так і алгоритми дерева рішень показали багатообіцяючі результати в прогнозуванні раку. Ці методи машинного навчання використовують потужність аналізу даних і розпізнавання шаблонів, щоб ідентифікувати шаблони та робити точні прогнози.

По-друге, продуктивність моделей логістичної регресії та дерева рішень може відрізнятися залежно від конкретного набору даних і характеру прогнозованого раку. У той час як логістична регресія є широко використовуваним алгоритмом, який можна інтерпретувати, який припускає лінійний зв'язок між предикторами та результатом, дерева рішень є нелінійними та можуть фіксувати більш складні взаємодії між змінними. Тому вибір алгоритму повинен базуватися на конкретних вимогах і особливостях завдання прогнозування раку.

Крім того, методи вибору ознак і попередньої обробки відіграють вирішальну роль у покращенні точності та надійності моделей прогнозування раку. Визначення відповідних функцій і зменшення розмірності може допомогти підвищити продуктивність моделі.

Загалом, застосування методів машинного навчання, зокрема алгоритмів логістичної регресії та дерева рішень, має великі перспективи для прогнозування раку. Ці методи є цінним інструментом для медичних працівників у виявленні осіб із групи ризику та уможливленні раннього втручання та персоналізованих стратегій лікування. Однак необхідні подальші дослідження та перевірка, щоб оптимізувати ці моделі, включити додаткові джерела даних і підвищити їх ефективність у реальних клінічних умовах.

Список використаних джерел:

1. Karthikeyan K., Vengatesan V., Venkatesh V., Prasanna G., Kumar S.S. Machine learning applications in cancer prognosis and prediction: A comprehensive review. *Computational and Structural Biotechnology Journal*. 2021. Vol. 19. 1533-1547. https://doi.org/10.1016/j.matpr.2021.03.625

2. Jenni A.M., Sidey-Gibbons C.J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*. 2019. Vol. 19. No. 64. https://doi.org/10.1186/s12874-019-0681-4

3. Kourou K., Exarchos T.P., Exarchos K.P. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*. 2015. Vol. 13. 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

4. Li Y., Hu G., Liu L., Xie Y., Xu Z. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Data in Brief*. 2021. Vol. 37. 100125. https://doi.org/10.1016/j.health.2022.100125

5. Sayed S. Machine Learning Is The Future Of Cancer Prediction. *Towards Data Science*. 2018. URL: https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa.

6. Azar A.S., Rikan S.B., Naemi A. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Medical Informatics and Decision Making*. 2022. Vol. 22. No. 68. https://doi.org/10.1186/s12911-022-02087-y

7. Jaber N. Can Artificial Intelligence Help See Cancer in New, and Better, Ways? *National Cancer Institute*. 2022. URL: https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging.

8. Mudawi N.A., Alazeb A. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors (Basel).* 2022. Vol. 22(11). 4132. https://doi.org/10.3390/s22114132

9. Kourou K., Exarchos T.P., Exarchos K.P. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal.* 2015. Vol. 13. 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

10. Yan-yan Y., Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*. 2015. Vol. 27. No. 2. 130-135. https://doi.org/10.11919/j.issn.1002-0829.215027.

References:

1. Karthikeyan, K., Vengatesan, V., Venkatesh, V., Prasanna, G., Kumar, S.S. (2021) Machine learning applications in cancer prognosis and prediction: A comprehensive review. *Computational and Structural Biotechnology Journal,* (19), 1533-1547. https://doi.org/10.1016/j.matpr.2021.03.625

2. Jenni, A.M., Sidey-Gibbons, C.J. (2019) Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology,* (19(64). https://doi.org/10.1186/s12874-019-0681-4

3. Kourou, K., Exarchos, T.P., Exarchos, K.P. (2015) Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal,* (13), 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

4. Li Y., Hu G., Liu L., Xie Y., Xu Z. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Data in Brief*. 2021. Vol. 37. 100125. https://doi.org/10.1016/j.health.2022.100125

5. Sayed S. Machine Learning Is The Future Of Cancer Prediction. *Towards Data Science*. 2018. URL: https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa.

6. Azar A.S., Rikan S.B., Naemi A. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Medical Informatics and Decision Making*. 2022. Vol. 22. No. 68. https://doi.org/10.1186/s12911-022-02087-y

7. Jaber N. Can Artificial Intelligence Help See Cancer in New, and Better, Ways? *National Cancer Institute*. 2022. URL: https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging.

8. Mudawi N.A., Alazeb A. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors (Basel).* 2022. Vol. 22(11). 4132. https://doi.org/10.3390/s22114132

9. Kourou K., Exarchos T.P., Exarchos K.P. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*. 2015. Vol. 13. 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

10. Yan-yan Y., Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*. 2015. Vol. 27. No. 2. 130-135. https://doi.org/10.11919/j.issn.1002-0829.215027.