ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЗНАЧЕННЯ ЙМОВІРНОСТІ ЗАХВОРЮВАННЯ НА ОСНОВІ ДІАГНОСТИЧНИХ ХАРАКТЕРИСТИК ПАЦІЄНТА

«Медична діагностика»

2020

АНОТАЦІЯ

Актуальність теми дослідження. Використання в медичних установах сучасних інформаційних комп'ютерних технологій дозволяє не тільки підвищити ефективність процесів обробки медичної інформації, а й відкриває великі перспективи у вирішенні багатьох проблем сучасної медицини. До теперішнього моменту створення і використання медичних систем підтримки прийняття рішень виділилося в окремий напрямок наукових досліджень, яке об'єднує результати, отримані для інтелектуальних систем, зокрема експертних систем, підходи Data Mining, а також сучасні концепції інформаційних технологій .

Мета роботи: проаналізувати існуючі моделі та методи визначення ймовірності захворювання на основі діагностичних характеристик пацієнта. Розробити програмне забезпечення для визначення ймовірності захворювання на основі удосконалених методів.

Об’єктом дослідження є процес визначення ймовірності захворювання на підставі діагностичних характеристик пацієнта.

Предметом дослідження є математичні моделі та методи для визначення ймовірності захворювання на підставі діагностичних характеристик пацієнта.

Методи дослідження: є розробка інформаційної системи для визначення ймовірності захворювання пацієнта із заданими діагностичними характеристиками на основі методів Data Mining. Для розробки програмного комплексу мова програмування C#.

Отримані результати: Були проаналізовані існуючі моделі та методи визначення ймовірності захворювання на основі діагностичних характеристик пацієнта. За допомогою мови програмування C# розроблено програмний комплекс, визначення ймовірності захворювання на основі діагностичних характеристик пацієнта.

Ключові слова: Data Mining, логістична регресія, Метод Байеса, Метод максимальної правдоподібності, ймовірність захворювання.

ЗМІСТ

ВСТУП

В реаліях сучасного світу, коли кожна людина має доступ до більшості благ, актуальність охорони здоров'я важко переоцінити. Контроль здоров'я людини − перш за все, захист від ризиків, як себе, так і оточуючих. Велика увага приділяється розробці ефективних і швидких механізмів встановлення діагнозу, пошуку найбільш актуальних відповідних методів і моделей, що дозволяють врахувати, як ризики, так і невизначеності, які можуть бути в людському житті і здоров'ї.

Одними з найбільш серйозних і поширених проблем зі здоров'ям є проблеми з серцем. Встановлення діагнозу являє собою складний механізм, який використовує теорію ймовірності. Моделі постановки діагнозу повинні враховувати величезну кількість чинників, оскільки визначають не тільки проблеми одного пацієнта, а й проблему в державі. Тому, методи статистичного аналізу не завжди дозволяють дати оцінку, стану пацієнта, в повній мірі. Через це, актуальним є застосування імовірнісних моделей і методів.

Якщо характеризувати стан цієї галузі, то насправді, можна сказати, що існує величезний розрив між рішеннями в теорії і на практиці.

Існує безліч літератури з приводу цієї проблеми, але сам механізм встановлення діагнозу реалізований слабко. З цієї причини практично відсутні продукти, які б вирішували проблему. Існують аналоги, але автори ретельно приховують механізми і моделі, які були закладені в програми, крім того, це програмне забезпечення коштує надто дорого.

У зв'язку з цим, актуальним завданням є створення інформаційної системи, яка буде визначати ймовірність захворювання.

Застосування даної розробки на практиці дозволить прискорити процес і якість діагностування пацієнтів з заданими наборами характеристик.

РОЗДІЛ 1

1.1 Актуальність задачі

Розвиток системи охорони здоров'я і вихід України на європейський рівень в медицині відбувається досить повільно. Відбувається це через цілу низку проблем.

Основна проблема − це брак фінансування, що неминуче відбивається на рівні медичної техніки та лабораторного обладнання, використовуваних технологій і методів діагностики. Актуальною проблемою охорони здоров'я є тенденція до збільшення ризику різних захворювань, до підвищення рівня смертності, що пов'язано з погіршеною екологією, низьким рівнем життя, несвоєчасною діагностикою, а також недостатньою профілактикою різних порушень здоров'я.

Використання в медичних установах сучасних інформаційних комп'ютерних технологій дозволяє не тільки підвищити ефективність процесів обробки медичної інформації, а й відкриває великі перспективи у вирішенні багатьох проблем сучасної медицини. До теперішнього моменту створення і використання медичних систем підтримки прийняття рішень (СППР) виділилося в окремий напрямок наукових досліджень [1], яке об'єднує результати, отримані для інтелектуальних систем, зокрема експертних систем (ЕС) [2], підходи Data Mining [3], а також сучасні концепції інформаційних технологій [4].

Одна з основних задач, вирішенню якої присвячена велика кількість публікацій, пов'язана зі створенням систем для медичної діагностики.

Обробці медичної інформації та подальшого прийняття рішень присвячені, наприклад, такі роботи: [1-6]. Незважаючи на це, ще досить багато питань у цій галузі залишилося відкритими, що відзначають, зокрема, автори [7-9].

1.2 Методи обробки медичної інформації

Формально задачу медичного діагностування можна представити як задачу класифікації, яка полягає в тому, щоб поставити у відповідність сукупності вхідних параметрів конкретне захворювання [10]. Основні підходи, які застосовуються для вирішення завдання медичної діагностики можна згрупувати наступним чином:

- логічний підхід;
- статистичний підхід;
- біонічний підхід.

Логічний підхід під час прийняття рішень в медицині є досить поширеним, тому що є прямим відображенням рішень лікаря [11]. Рішення лікаря під час діагностичного процесу повинні бути впевненими, послідовними та мати обґрунтування.

До групи статистичних методів можна віднести байесовський підхід [12], методи дискримінантного аналізу [13], вивід, заснований на прецедентах [14]. Використання теореми Байеса при визначенні класу захворювання − досить поширений підхід через свою простоту, наочність і застосування простих математичних обчислень. Дискримінантний аналіз характеризується наявністю великої кількості обчислень, виявленням різного роду зв'язків між ознаками, з'ясуванням впливу останніх на результат діагностування, що зазвичай тягне до труднощів при вирішенні поставленого медичного завдання [15].

Біонічний підхід являє собою процес штучного відтворення тих структур і процесів, які характерні людському мозку. В рамках цього підходу для вирішення завдання діагностування було розроблено нейромережевий підхід [16]. Переваги, властиві цьому підходу досить великі:

- здатність до адаптації (навчання і самонавчання);
- паралельність обробки інформації;
- робасність (стійкість до окремих збоїв).

Нейронні мережі (НМ) використовують при вирішенні завдань медичної діагностики, завдань класифікації, кластеризації, апроксимації, прогнозування [17].

Таким чином, завдання медичної діагностики являє собою досить складну задачу з огляду на те, що дані про пацієнта є не достатньо структурованими і мають різний характер. Частина необхідної інформації, що стосується пацієнта, зазвичай відсутня, що вносить додаткові труднощі під час обробки медичних даних; частина інформації носить якісний характер, бо її визначає лікар, тобто присутня частка суб'єктивізму; частина інформації відображає результати аналізів, що говорить про необхідність врахування фактору випадковості через помилки вимірювань. На сьогоднішній день склалося кілька підходів роботи з невизначеністю в задачах медичного діагностування. Імовірнісний підхід являє собою підхід, коли невідомі фактори статистично стійкі і тому являють собою випадкові величини або випадкові події [18]. При цьому повинні бути визначені всі необхідні статистичні характеристики: закони розподілу і їх параметри, функції або щільності розподілу ймовірностей, що, в свою чергу, вносить додаткові труднощі.

Завдання медичної діагностики – складне і багатогранне завдання, рішенням якого займалося не одне покоління вчених [19]. На сьогоднішній день розроблено велику кількість підходів до її вирішення, але кожен з них має свої недоліки і обмеження, тому необхідність розробки ефективних моделей і методів медичної діагностики не втратила своєї актуальності.

1.3 Data mining

Data mining – збірна назва, що використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності.

Основу методів data mining складають всілякі методи класифікації, моделювання і прогнозування, засновані на застосуванні дерев рішень, штучних нейронних мереж, генетичних алгоритмів, еволюційного програмування, асоціативної пам'яті, нечіткої логіки. До методів data mining нерідко відносять статистичні методи (дескриптивний аналіз, кореляційний і регресійний аналіз, факторний аналіз, дисперсійний аналіз, компонентний аналіз, дискримінантний аналіз, аналіз часових рядів, аналіз виживаності, аналіз зв'язків).

Одне з найважливіших призначень методів data mining полягає в наочному поданні результатів обчислень (візуалізація), що дозволяє використовувати інструментарій data mining людьми, які не мають спеціальної математичної підготовки.

Data mining використовуються системи управління базами даних, статистичні методи аналізу і методи штучного інтелекту. Завдання, які вирішуються методами data mining, прийнято розділяти на описові і ті, які передбачають.

Для задач класифікації характерне «навчання з учителем», при якому побудова (навчання) моделі проводиться за вибіркою, що містить вхідні та вихідні вектори.

Для задач кластеризації та асоціації застосовується «навчання без учителя», під час якого побудова моделі проводиться за вибіркою, в якій немає вихідного параметра. Значення вихідного параметра підбирається автоматично в процесі навчання.

Для завдань скорочення опису характерна відсутність поділу на вхідні і вихідні вектори.

1.4 Постановка задачі

Задачі дослідження наступні:

– проаналізувати існуючі моделі та методи визначення ймовірності захворювання на основі діагностичних характеристик пацієнта;

– розробити програмне забезпечення для визначення ймовірності захворювання на основі вибраних методів.

Задача дослідження полягає у визначенні, аналізі та реалізації моделей та методів, які дозволяють оцінити ймовірність захворювання пацієнта із заданими діагностичними характеристиками.

Об'єктом дослідження в даній роботі є процес визначення ймовірності захворювання на підставі діагностичних характеристик пацієнта.

Предметом дослідження є математичні моделі та методи для визначення ймовірності захворювання на підставі діагностичних характеристик пацієнта.

Метою дослідження є розробка інформаційної системи для визначення ймовірності захворювання пацієнта із заданими діагностичними характеристиками на основі методів Data Mining.

РОЗДІЛ 2
МОДЕЛІ ТА МЕТОДИ ВИЗНАЧЕННЯ ІМОВІРНОСТІ ЗАХВОРЮВАННЯ НА ОСНОВІ ДІАГНОСТИЧНИХ ХАРАКТЕРИСТИК ПАЦІЄНТА

2.1 Логістична регресія

У статистиці, логістична модель [20] широко використовується як статистична модель, яка в своїй основній формі використовує логістичну функцію, щоб моделювати бінарну залежну змінну. У логістичній моделі, лог-коефіцієнти для значення з написом «1» представляє собою лінійну комбінацію з одного або декількох незалежних змінних; кожна незалежна змінна може бути бінарною змінною (два класи, закодовані індикаторною змінною) або безперервною змінною (будь-яке дійсне значення). Відповідна ймовірність-значення, позначене «p», може варіюватися від 0 (безумовно значення «0») до 1 (безумовно значення «1»); функція, яка перетворює лог-шанси в ймовірність, є логістичною функцією $p' = \ln\left(\frac{p}{1+p}\right), де\ p = \frac{e^x}{1+e^x}$ (рис 1.1), звідси і назва.

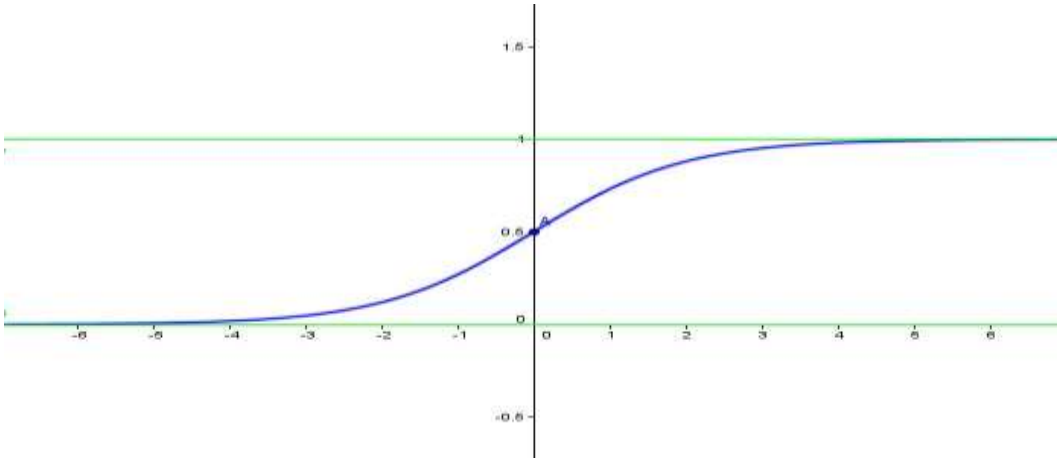

Рисунок 1.1 – Логістична крива

2.2 Оцінювання параметрів логістичної регресії на основі методу оцінки шансів та імовірностей

Оцінювання шансів (ОШ) [21] − статистичний показник, один з основних способів описати в чисельному вираженні те, наскільки відсутність або наявність певного результату пов'язана з присутністю чи відсутністю певного фактора в конкретній статистичній групі.

У науковій медичній літературі показник відношення шансів було вперше згаданий в 1951 році в роботі Дж. Корнфілда. Згодом даним дослідником були опубліковані роботи, в яких наголошувалося на необхідності розрахунку 95% довірчого інтервалу для співвідношення шансів.

Співвідношення шансів дозволяє оцінити зв'язок між певним результатом і фактором ризику та дозволяє порівняти групи досліджуваних за частотою виявлення певного фактора ризику. Важливо, що результатом застосування співвідношення шансів є не тільки визначення статистичної значущості зв'язку між фактором і результатом, але і її кількісна оцінка.

Оцінка шансів − це значення дробу, в чисельнику якого, знаходяться шанси певної події для першої групи, а в знаменнику шанси тієї ж події для другої групи.

Зручним способом є розрахунок співвідношення шансів зі зведенням даних в таблицю 2 на 2, як в таблиці 2.1:

Таблиця *2.1 – Приклад таблиці для методу оцінки шансів*

	Результат є (1)	Результату немає (0)	Всього
Фактор ризику є (1)	A	B	A+B
Фактора ризику немає (0)	C	D	C+D
Всього	A+C	B+D	A+B+C+D

Для даної таблиці співвідношення шансів розраховується за такою формулою:

$$OR = \frac{A \cdot D}{B \cdot C}. \tag{2.1}$$

Дуже важливо дати оцінку статистичній значущості виявленого зв'язку між результатом і фактором ризику. Пов'язано це з тим, що навіть коли значення співвідношення шансів невисокі, близькі до одиниці, зв'язок може виявитися істотним і має бути врахованим у статистичних висновках. І навпаки, коли значення ОШ великі, показник виявляється статистично незначним, і, отже, виявленим зв'язком можна знехтувати.

Для оцінки важливості відношення шансів розраховуються для кордону 95% довірчого інтервалу (ДІ). Формула для знаходження значення верхньої межі 95% ДІ:

$$e^{\ln(OR)+1.96\cdot\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}}. \tag{2.2}$$

Формула для знаходження значення нажньої межі 95% ДІ:

$$e^{\ln(OR)-1.96\cdot\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}}. \tag{2.3}$$

Якщо співвідношення шансів перевищує одиницю, то це означає, що шанси виявити фактор ризику більше в групі з наявністю результату. Співвідношення шансів, що має значення менше одиниці, свідчить про те, що шанси виявити фактор ризику більше в другій групі.

2.3 Метод Байеса для оцінки імовірності захворювання

У теорії ймовірностей і статистиці, теорема Байеса для оцінки імовірності захворювання описує ймовірність як події, на основі попереднього знання умов, які можуть бути пов'язані з подією. Одним з багатьох застосувань теореми Байеса для оцінки імовірності захворювання є байесовский висновок, особливий підхід до статистичного висновку. При застосуванні ймовірності, включені в теорему Байеса для оцінки імовірності захворювання, можуть мати різні імовірнісні інтерпретації [22].

Є кілька форм формули Байеса для оцінки імовірності захворювання, для подій A і B за умови, де $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.4}$$

$$P(A|B) \propto P(B|A)P(A), \tag{2.5}$$

що означає, пропорційність A для даного B.

Якщо події $A_1, A_2, \ldots$, взаємовиключні та вичерпні, тобто можлива тільки одна з подій, одночасно дві події не можуть статися разом, ми можемо визначити коефіцієнт пропорційності, орієнтуючись на те, що їх ймовірності в сумі повинні складати одиницю.

$$P(A|B) = c \cdot P(A) \cdot P(B|A), \tag{2.6}$$

$$P(\neg A|B) = c \cdot P(\neg A) \cdot P(B|\neg A). \tag{2.7}$$

Об'єднавши ці дві формули, ми отримаємо:

$$c = \frac{1}{P(A)\cdot P(B|A) + P(\neg A)\cdot P(B|\neg A)} = \frac{1}{P(B)}. \quad (2.8)$$

Часто простір подій (таких як A_j) визначено в термінах $P(A_j)$ і $P(B|A_j)$. Саме в цьому випадку корисно визначити $P(B)$, застосувавши формулу повної ймовірності:

$$P(B) = \sum_j P(B|A_j)P(A_j) => P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}. \quad (2.9)$$

Якщо A є бінарною змінною:

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}. \quad (2.10)$$

2.4 Метод максимальної правдоподібності для оцінки параметрів логістичної регресії

Нехай Y − реалізація N-мірної випадкової величини, розподіленої як:

а) $P_\theta(x)$ (ймовірність) − в разі дискретного розподілу.

б) $p_\theta(x)$ (щільність) − в разі безперервного розподілу.

Тут $P_\theta(x)$ ($p_\theta(x)$) характеризує сімейство розподілів, задається параметром $\theta \in \Theta$, $\Theta \subset R^m$ − простір параметрів. В економетрії прийнято казати про це сімейство розподілів, як про те, яке породжує дані процесу.

Функція $L(Y,\theta) = P_\theta(Y)$ (відповідно $L(Y,\theta) = p_\theta(Y)$) називається функцією правдоподібності.

Оцінкою максимальної правдоподібності (МП) $\hat{\theta}$, називається рішення задачі:

$$L(Y,\theta) \to max_{\theta \in \Theta}. \tag{2.11}$$

Такий метод оцінювання називають методом максимальної правдоподібності [23]. Зазвичай зручніше користуватися логарифмічною функцією максимальної правдоподібності:

$$l(Y,\theta) = \ln\left(L(Y,\theta)\right). \tag{2.12}$$

В окремому випадку вектор спостережень являє собою вибірку незалежних однаково розподілених випадкових величин: $Y_i \sim HOP, i = 1 \ldots N$.

При цьому

$$L(Y,\theta) = \prod_i L_i(Y_i,\theta), \tag{2.13}$$

$$l(Y,\theta) = \prod_i l_i(Y_i,\theta). \tag{2.14}$$

Вектор спостережень Y складається з залежних між собою і/або неоднаково розподілених випадкових величин, тому не є вибіркою в звичайному сенсі слова. У загальному випадку ця рівність теж буде вірна, якщо позначити:

$$L_i(Y_i,\theta) = p_\theta(Y_i \mid Y_{i-1},\ldots,Y_1) \tag{2.15}$$

$$l_i(Y_i,\theta) = \ln\left(L_i(Y_i,\theta)\right) \tag{2.16}$$

Тим самим задається розбиття функції правдоподібності на вклади окремих спостережень.

Оцінка максимальної правдоподібності є функцією вектору спостережень: $\hat{\theta} = \hat{\theta}(Y)$, тому це теж випадкова величина.

Нехай функція правдоподібності диференційована по θ і досягає максимуму у внутрішній точці ($\hat{\theta} \in \text{int}(\Theta)$), тоді оцінка МП $\hat{\theta}$ повинна задовольняти наступним умовам першого порядку:

$$\frac{\partial L}{\partial \theta}\left(Y, \hat{\theta}\right) = 0 \tag{2.17}$$

$$\frac{\partial l}{\partial \theta}\left(Y, \hat{\theta}\right) = 0 \tag{2.18}$$

Таким чином, градієнт логарифмічної функції правдоподібності g(θ) при $\theta = \hat{\theta}$ має дорівнювати нулю.

Для того, щоб оцінки, що задовольняють цим рівнянням правдоподібності, дійсно давали максимум правдоподібності, необхідно і достатньо, щоб були виконані умови другого порядку (припускаємо, що функція правдоподібності двічі диференційована). А саме, матриця Гессе («гессіан») логарифмічною функції правдоподібності повинна бути всюди негативно визначена. Матриця Гессе H за визначенням є матрицею других похідних:

$$H_{ij}\left(Y, \theta\right) = \frac{\partial^2 l}{\partial \theta_j \partial \theta_i}\left(Y, \theta\right), \tag{2.19}$$

де $i, j = 1 \ldots m.$

За допомогою матричного диференціювання можна записати «гессіан» у вигляді:

$$H = \frac{\partial^2 l}{\partial \theta \partial \theta^T}. \tag{2.20}$$

У деяких моделях функція правдоподібності необмежена зверху і не існує оцінок максимальної правдоподібності в сенсі, наведеного вище визначення. Відповідно до альтернативного визначення оцінками максимальної

правдоподібності називають рішення рівняння правдоподібності, є локальними максимумами функції правдоподібності, рішенням рівняння правдоподібності.

2.5 Порівняння методів

Таблиця *2.2 – Порівняння методів*

Назва метода	Рек. об’єм вибірки	Переваги	Недоліки
Оцінювання параметрів логістичної регресії на основі методу оцінки шансів	Будь-який	Працює із будь-яким об’ємом вибірки	Дає досить неінформатив-ний результат
Метод імовірностей для оцінювання параметрів логістичної регресії	Будь-який	Працює із будь-яким об’ємом вибірки	Дає досить неінформатив-ний результат
Метод Байеса для оцінки імовірності захворювання	До 3000	Показує високу точність на малій виборці	Зі збільшенням об’єму вибірки погіршується точність
Метод максимальної правдоподібності для оцінки параметрів логістичної регрессії	Більше ніж 3000	На великих вибірках показує велику точність	Для малих вибірок існують методи із більшою точністю

РОЗДІЛ 3
АЛГОРИТМІЧНІ МОДЕЛІ ВИЗНАЧЕННЯ ІМОВІРНОСТІ ЗАХВОРЮВАННЯ ПАЦІЄНТА

3.1 Вхідні дані та вихідні дані

Формат даних представлений в табл. 3.1. Вхідні і вихідні дані для всіх методів однакові.

Таблиця *3.1 – Вхідні дані*

Назва	Тип	Інтервал
«Вік»	Ціле число	[0-90]
«Стать»	Логічний тип	0,1
«Тип болю в грудях»	Ціле число	[0-4]
«Верхній тиск»	Ціле число	[100-180]
«Холестерол»	Ціле число	[100-500]
«Цукор в крові більше 120 одиниць»	Логічний тип	0,1
«Характеристика ЕКГ»	Ціле число	[0-4]
«Пульс»	Ціле число	[80-220]
«Ангіна»	Логічний тип	0,1
«Peak»	Дрібне число	0,0.1,…,4
«Slope»	Ціле число	[0-4]
«Colored vessels»	Ціле число	[0-3]
«Thal»	Ціле число	[0-3]
«Class»	Логічний тип	0,1

Інші параметри для функціонування системи генеруються або розраховуються автоматично, в програмному забезпеченні.

В результаті роботи програми ймовірність захворювання розраховується для кожного пацієнта окремо, це означає, що методи повертають саму ймовірність, а не логіт-регресію. Єдиним вихідним параметром є ймовірність належності до того чи іншого класу (табл. 3.2).

Таблиця *3.2 – Вихідні дані*

Назва	Тип	Інтервал
Ймовірність захворювання	Дрібне число	[0.0-1.0]

3.2 Алгоритмічна модель оцінювання параметрів логістичної регресії на основі метода оцінки шансів та імовірностей

Алгоритмічна модель оцінювання параметрів логістичної регресії на основі метода оцінки шансів полягає в наступному:

Етап 1. Формування чотирьохполих таблиць з вхідних даних.

Етап 2. Обчислення шансів потрапити в групу «хворий» для всіх категорій:

$$Ch_{y=1,C_n} = \frac{Count(y=1,C_n)}{Count(y=0,C_n)}, \tag{3.1}$$

де $Count(y=q,C_n)$ – кількість пацієнтів з результатом $y=q$ для n-ої категорії.

Етап 3. Обчислення відношення шансів:

$$OR(\frac{C_n}{C_m}) = \frac{Ch_{y=1,C_n}}{Ch_{y=1,C_m}}. \tag{3.2}$$

Етап 4. Оцінка отриманих результатів.

Алгоритмічна модель метода імовірностей для оцінювання параметрів логістичної регресії полягає в наступному:

Eman 1. Обчислення експериментальної ймовірності захворювання:

$$c_{\exp(n)} = \frac{Count(y=1, C_n)}{Count(y=1)}, \tag{3.3}$$

де $Count(y=q, C_n)$ – кількість пацієнтів з результатом $y=q$ для n-ої категорії,

$Count(y=q)$ – кількість пацієнтів з результатом $y=q$ для всіх категорій.

Eman 2. Розрахунок β для заданого значення пояснювальної змінної $P(y=1|x)=\rho(x)$, де $\rho(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$, $g(x)=\beta_0+\beta_1 C_1+...+\beta_n C_n$, n – кількість класів:

$$\beta_n = \ln\left(\frac{c_{\exp(n)}}{1-c_{\exp(n)}}\right) - \sum_{m=0}^{n-1}\beta_m, \tag{3.4}$$

змінні квантування використовуємо у вигляді бінарного числа досліджуваної групи (по рахунку).

Етап 3. Обчислення результату:

$$P(y=1|C_n) = \frac{e^{\sum_{i=0}^{n}\beta_i C_i}}{1-e^{\sum_{i=0}^{n}\beta_i C_i}}. \tag{3.5}$$

Eman 4. Оцінка отриманих результатів.

3.3 Алгоритмічна модель методу Байеса

Алгоритм методу Байеса для оцінки імовірності захворювання полягає в наступному (в разі відомої інформації про низьку кореляційну залежність перейти до етапу 3):

Етап 1. Шукаємо інтервал коефіцієнтів парної кореляції:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}, \quad (3.6)$$

де n – число спостережень,

x_i – i-е спостережуване значення незалежної випадкової величини,

y_i – i-е спостережуване значення залежної випадкової величини,

r_{xy} – коефіцієнт парної кореляції.

Етап 2. Оцінка інтервалів коефіцієнтів [a, b]:

а) max (| a |, | b |)> 0.4 – дані не слабо корельовані, дослідження не може бути продовжено.

б) max (| a |, | b |) <0.4 – перехід до етапу 3.

Етап 3. Знаходження апріорної ймовірності для кожного класу:

$$P(y = n) = \frac{Count(y = n)}{Count(\sum_{i=0}^{n} y = i)}, \quad (3.7)$$

де $Count(\sum_{i=0}^{n} y = i)$ – кількість всіх пацієнтів,

$Count(y = q)$ – кількість пацієнтів з результатом $y = q$ для всіх категорій.

Етап 4. Знаходження умовних ймовірностей для конкретного випадку:

$$P(X \mid C_n) = \prod_{i=0}^{n} P(X_i \mid C_n), \quad (2.8)$$

де C_n – вхідні параметри досліджуваного пацієнта.

Етап 5. Знаходження ймовірності захворювання конкретного пацієнта:

$$p = P(X \mid C_n)P(C_n), \tag{2.9}$$

де C_n – вхідні параметри досліджуваного пацієнта.

Етап 6. Оцінка отриманих результатів.

3.4 Алгоритмічна модель методу максимальної правдоподібності

Алгоритм методу максимальної правдоподібності полягає в наступному:

Етап 1. Визначення розподілу випадкової величини:

на даному етапі необхідно порівнювати графіки випадкових величин і знаходити кореляцію між даними так само як і зважену суму квадратів відхилень. Чим ближче кореляція до 1 і менше зважена сума квадратів відхилень − тим краще.

Етап 2. Визначення сумарної ймовірності вибірки:

$$P(X^n, \theta) = \prod_{i=1}^{n} F(X^n, \theta). \tag{3.10}$$

де $F(X^n, \theta)$ – функція розподілу випадкової величини з невідомим параметром θ.

Етап 3. Максимізація сумарної ймовірності вибірки по параметру θ:

$$\hat{\theta}_{ОМП} = \arg\max_{\theta} \ln P(X^n, \theta), \tag{3.11}$$

де $\ln P(X^n, \theta) = \ln(\prod_{i=1}^{n} F(X^n, \theta))$.

Для цього треба вирішити:

$$\frac{\partial P(X^n, \theta)}{\partial \theta} = 0. \tag{3.12}$$

Етап 4. Оцінка отриманих результатів.

РОЗДІЛ 4
ОПИС РЕЗУЛЬТАТІВ РОБОТИ ІНФОРМАЦІЙНОЇ СИСТЕМИ

4.1 Опис програмного продукту

Інтерфейс розробленої програми інтуїтивно зрозумілий користувачеві, бо при розробці використовувалися стандартні елементи управління C#, а так само, вони були максимально описані.

На рисунку 4.1 зображений інтерфейс системи.

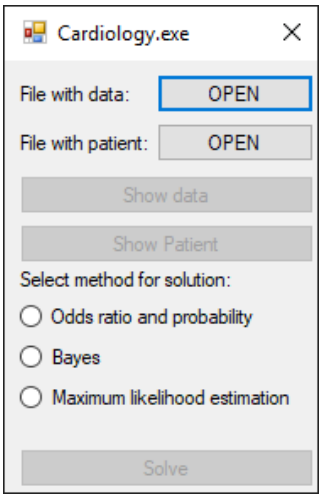

Рисунок 4.1 – Інтерфейс системи

На головному екрані форми, можна бачити 4 блоки:

– завантаження даних;

– демонстрація завантажених даних;

– вибір методу, яким знаходити ймовірність;

– рішення проблеми.

Для доступу до решти блоків потрібно завантажити дані для роботи в форматі «*.csv». Було вирішено відмовитися від зміни даних через форму додатка через те, що такий функціонал здався надлишковим.

Після того, як дані були завантажені, можна перейти до блоку демонстрації, цей блок ми можемо бачити на рисунку 4.2.

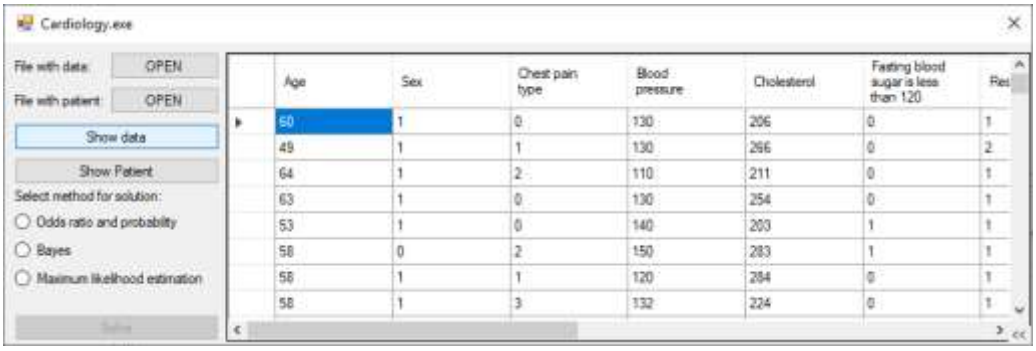

Рисунок 4.2 – Демонстрація даних

Після необов'язкового перегляду даних слід обрати метод для визначення ймовірності захворювання пацієнта, з урахуванням поточних характеристик, для цього в блоці «Select method for solution» необхідно обрати один з трьох представлених варіантів, після чого буде доступний блок рішення (рис 4.3).

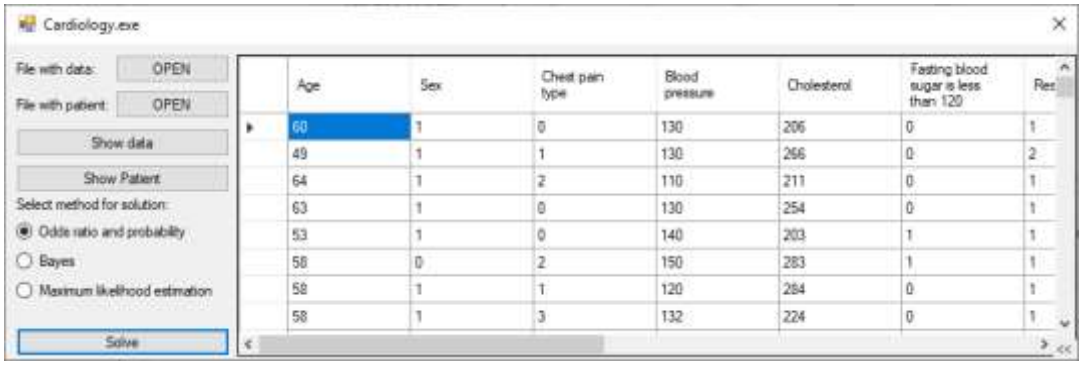

Рисунок 4.3 – Вибір методу

Після чого, в залежності від обраного методу, відкриється нова форма. Якщо вибрати пункт «Odds ratio and probability», то результат можна побачити на рисунку 4.4.

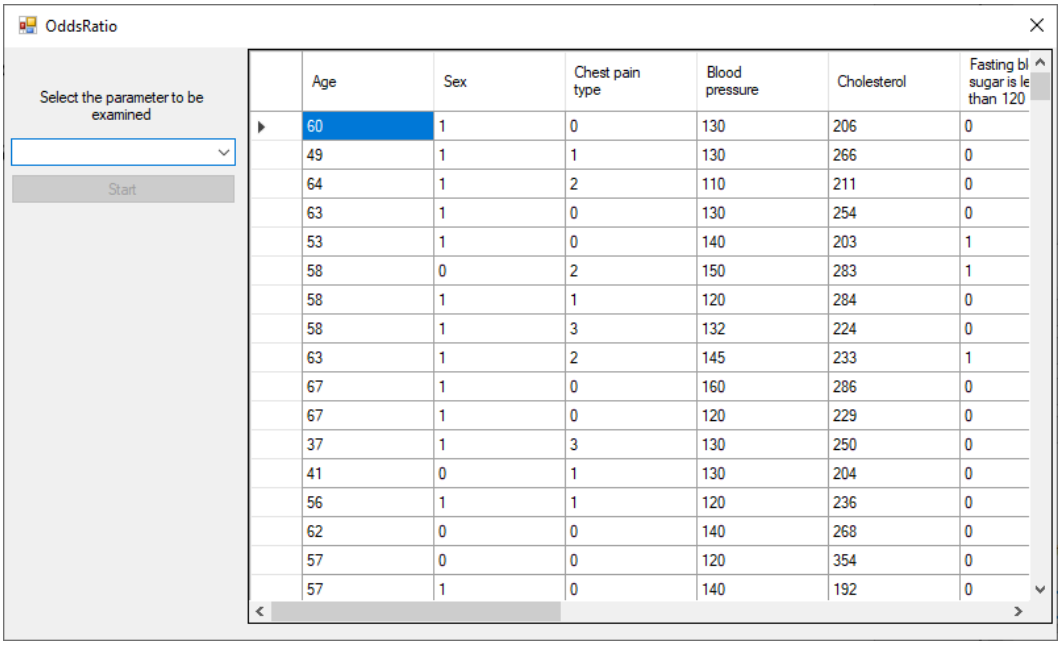

Рисунок 4.4 – Форма методу для оцінки параметрів логістичної регресії на основі методу оцінки шансів і ймовірностей

На даній формі можна відразу побачити набір даних, з яким ми працюємо й обрати параметр, за яким і буде визначатися ймовірність.

4.2 Результати, отримані при роботі з програмою

Розглянемо роботу програмного забезпечення на прикладі вхідних даних, які були згенеровані випадковим чином.

Розглянемо результати методів, на прикладі вхідних даних (табл. 4.1), отриманих в результаті роботи програмного забезпечення.

Проаналізуємо метод імовірностей для оцінювання параметрів логістичної регресії. Параметром дослідження для даного методу була обрана «Характеристика ЕКГ». Для початку складемо таблицю для даного параметра (табл. 4.2)

Таблиця *4.1 – Вхідні дані*

Назва	Вхідне значення
«Вік»	61
«Стать»	Male
«Тип болю в грудях»	NoTang
«Верхній тиск»	150
«Холестерол»	243
«Цукор в крові більше 120 одиниць»	True
«Характеристика ЕКГ»	Normal
«Пульс»	137
«Ангіна»	True
«Peak»	1
«Slope»	Flat
«Colored vessels»	0
«Thal»	Normal

Таблиця *4.2* – Дані про пацієнтів станом ЕКГ

Результат	Normal	Нур	Abnormal	Всього
y=0	96	68	1	165
y=1	56	79	3	138
Всього	152	147	4	303

Далі у $\rho(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$ для функції $g(x)=\beta_0+\beta_1 C_1+\beta_2 C_2$ визначимо змінні квантування. У рядку з класом «Normal» змінні квантування будуть рівні: $C_1=C_2=0$. У рядку з класом «Нур» $C_1=1, C_2=0$. У рядку з класом «Abnormal» $C_1=C_2=1$.

Для вхідних даних функція виглядає як $g(x)=\beta_0$

Далі шукаємо експериментальні ймовірності і спільно з цим шукаємо коефіцієнти β.

Експериментальну ймовірність захворювання для категорії «Normal» можна знайти, поділивши число позитивних результатів на загальну кількість випадків

$$c_{exp} = 56 / 152 = 0.37. \tag{4.1}$$

Звідси β_0 може бути знайдений як:

$$\beta_0 = \ln\left(\frac{c_{exp}}{1 - c_{exp}}\right) = -0.532. \tag{4.2}$$

Подальші обчислення не потрібні.

Імовірність того, що вихідна змінна y буде дорівнює одиниці (тобто пацієнт буде хворий) для категорії «Normal» розраховується за формулою:

$$P(y = 1 | x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{e^{-0,532}}{1 + e^{-0,532}} \approx 0.37. \tag{4.3}$$

Або $P(y = 0 | x) = 1 - P(y = 1 | x) \approx 0.63$

Результат роботи методів, в ході вивчення, був порівняний з результатами роботи програми SPSS Statistics, на основі результатів можна сказати, що модель адекватно описує дану сукупність.

ВИСНОВКИ

Було розглянуто проблеми та основні підходи до вирішення задач медичної діагностики.

Були проаналізовані існуючі моделі та методи визначення ймовірності захворювання на основі діагностичних характеристик пацієнта.

В результат дослідження була отримана інформаційна система для визначення ймовірності захворювання на основі діагностичних характеристик пацієнта. Були вирішені всі завдання, поставлені на початку дослідження. У роботі наведені основні моделі та методи для роботи з представленим набором даних. Для розрахунку імовірності захворювання було обрано логістичну регресію. Дана модель дозволяє отримувати значення-результати експрес-діагностування для кожного пацієнта в будь-який момент часу.

На підставі методів визначення оцінки параметрів логістичної регресії були розроблені алгоритмічні моделі. Представлені конкретні приклади використання описаних алгоритмічних моделей.

Наведено критерії якості діагностичних моделей та проведена перевірка якості отриманої моделі за критеріями: Хосмера-Лемешова, R-квадрат Нейджелкерка, R-квадрат Кокса та Снелла.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mark, D. B. Chapter 3. Decision-making in clinical medicine. In A. S. Fauci, E. Braunwald, D. L. Kasper, S. L. Hauser, D. L. Longo, J. L. Jameson, & J. Loscalzo (Eds.), Harrison's principles of internal medicine (17th ed.). – New York: McGraw-Hill, 2008 – pp. 6-52.

2. Singla J., Grover D., Bhandari A. Medical Expert Systems for Diagnosis of Various Diseases // International Journal of Computer Applications, vol. 93. – pp. 36-43.

3. Stragier J, Vandewiele G, Coppens P, Ongenae F, Van den Broeck W, De Turck F, De Marez L Data Mining in the Development of Mobile Health Apps Assessing In-App Navigation Through Markov Chain Analysis // J Med Internet Res, vol. 21(6), 2019 - e11934.

4. Ahuja A.S. The impact of artificial intelligence in medicine on the future role of the physician. // Peer Journal, vol. 7, 2019, - e7702.

5. Дюк В. Обработка данных на ПК в примерах / В. Дюк. – СПб: Питер, 1997. – 240 с.

6. Круглов В.В. Искусственные нейронные сети. Теория и практика / В.В. Круглов, В.В. Борисов. – 2-е изд., стереотип. – М.: Горячая линия – Телеком, 2002. – 382 с.

7. Ластед Л. Введение в проблему принятия решений в медицине / Л. Ластед; пер. с англ. И.М. Быховского. – М.: Мир, 1971. – 283 с.

8. Мисюк Н.С. Диагностические алгоритмы / Н.С. Мисюк, А.М. Гурленя, В.В. Лозовик. – Мн.: Вышэйшая школа, 1970. – 187 с.

9. Мацуга О.М. Інформаційна технологія обробки неоднорідних медичних даних для підтримки прийняття рішень під час діагностики: дис. канд. техн. наук: 05.13.06 – Д.: Дніпропетровський національний ун-т, 2007. – 209 с.

10. Мельник К.В. Процедура диагностирования состояния сердечно-сосудистой системы пациента на основе нечеткой логики / К.В. Мельник, А.Е. Голоскоков // Вестник НТУ «ХПИ». – Х., 2008. – № 49. – С. 101-104.

11. Arocha J.F., Wang D., Patel V.L. Identifying reasoning strategies in medical decision making: A methodological guide // Journal of Biomedical Informatics, Vol. 38, Iss. 2, 2005 – pp. 154-171.

12. Verma V., Mishra A.K., Narang R. Application of Bayesian Analysis in Medical Diagnosis // Journal Practical Cardiovascular Science, vol. 5, 2019 – pp. 136-41.

13. Ren, Zh. Application of Discriminant Analysis in Medical Diagnosis // DEStech Transactions on Social Science, Education and Human Science. 2017, 8311.

14. Miotto R., Kidd B.A., Dudley J.T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records // Scientific Reports, vol. 6, no. 1, 2016 - p. 26094.

15. Ho C.W.L., Soon D., Caals K., Kapur J. Governance of automated image analysis and artificial intelligence analytics in healthcare // Clinical Radiology, Vol. 74, Iss. 5, 2019 – pp. 329-337.

16. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. // PLoS One, vol. 14(2), 2019 - e0212356.

17. Jiang F., Jiang Y., Zhi H., Dong Y., Li H., Ma S. Artificial intelligence in healthcare: past, present and future // Stroke Vasc Neurol, vol. 2(4), 2017 – pp.230–243.

18. Abdel-Badeeh M. Salem. Case Based Reasoning Technology for Medical Diagnosis / Abdel-Badeeh M. Salem // World Academy of Science, Engineering and Technology. – 2007. – № 31. – P. 9-13.

19. Rubin G, The expanding role of primary care in cancer control [Text] / Rubin G, Berendsen A, Crawford SM, Dommett R, Earle C. Lancet Oncol, 2015 – pg. 31–72.

20. Kuhle S., Maguire B., Zhang H. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. // BMC Pregnancy Childbirth, vol. 18(1), 2018 – pp. 333.

21. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. // Journal of Biomedical Informatics, vol. 35, 2002 – pp. 352–359.

22. Ashby D. Bayesian statistics in medicine: a 25 year review // Statistics in Medicine, vol. *25*, 2006 – pp. 3589-3631.

23. Schomaker M., Luque-Fernandez M.A., Leroy V., Davies M.A. Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions // Statistics in Medicine, vol. 38, iss. 24, 2019 – pp. 4888-4911