МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет імені В.Н.Каразіна
Факультет математики і інформатики
Кафедра прикладної математики

Кваліфікаційна робота
магістра

на тему «*Прогнозування ризику виникнення хвороб серця за допомогою методів штучного інтелекту та машинного навчання*»

Виконав: студент групи МП-61
2-го курсу
спеціальність 113 – прикладна математика
освітньо-наукова програма «Прикладна математика»
Продащук М. В.

Наукові керівники:
д.т.н., проф. Ромашов Ю.В.
Lead Statistical Programmer/Analyst, Intego Group
Невмержицька О. А.

Рецензент: *Statistical Programmer/Analyst, Intego Group*
Бугерчук Х. Ю.

Харків – 2024 рік

Анотація

Продащук М. В. «Прогнозування ризику виникнення хвороб серця за допомогою методів штучного інтелекту та машинного навчання»

Проаналізовані підходи у сфері штучного інтелекту та машинного навчання, які застосовуються в області медицини.

Розроблена та описана модель прогнозування ризику виникнення хвороб серця (зокрема, ішемічної хвороби серця) з використанням багатомірної логістичної регресії. Обрані інструменти реалізації моделі.

Відібрані дані для реалізації та аналізу розробленої моделі, проведена обробка даних.

Реалізована модель, а також проведений аналіз ефективності реалізованої моделі.

Ключові слова: ішемічна хвороба серця, прогнозування ризику, штучний інтелект, машинне навчання, логістична регресія, мова програмування Python, бібліотеки Pandas та Scikit-learn (sklearn), обробка даних.

Summary

Prodashchuk M.V. «Prediction of heart diseases risks with the help of artificial intelligence and machine learning methods»

The approaches in the field of artificial intelligence and machine learning applied in medicine have been analyzed.

A predictive model for assessing the risk of heart diseases (specifically ischemic heart disease) has been developed and described using multidimensional logistic regression.

The implementation tools for the model have been chosen. Data for the implementation and analysis of the developed model have been selected, and data processing has been conducted.

The model has been implemented, and an analysis of its effectiveness has been performed.

Keywords: ischemic heart disease, risk prediction, artificial intelligence, machine learning, logistic regression, Python programming language, Pandas and Scikit-learn (sklearn) libraries, data processing.

ЗМІСТ

ВСТУП

Ішемічна хвороба серця є однією з найпоширеніших та серйозних серцево-судинних захворювань у світі. Вона виникає внаслідок недостатнього постачання крові до серцевого м'язу через блокування артерій, що живлять серце. Ішемічна хвороба серця може призвести до серцевого нападу, стенокардії, серцевої недостатності та інших серйозних ускладнень, які значно погіршують якість життя та підвищують ризик смерті.

Важливо вчасно визначати ризик виникнення ішемічної хвороби серця, оскільки це дозволяє приймати передбачувані заходи для профілактики та лікування. Оцінка ризику включає аналіз клінічних показників, таких як вік, стать, куріння, артеріальний тиск, рівень холестерину та інші фактори, що сприяють розвитку серцево-судинних захворювань. Чим точніше визначається ризик, тим ефективніші можуть бути запроваджені профілактичні стратегії та вчасне лікування.

Новітні технології, зокрема, методи штучного інтелекту та машинного навчання, можуть значно полегшити процес визначення ризику виникнення ішемічної хвороби серця. Вони дозволяють аналізувати великі обсяги клінічних даних та виявляти складні зв'язки між різними факторами ризику, що дозволяє створювати індивідуалізовані та точні моделі для прогнозування ризику захворювання. Такий підхід є перспективним у поліпшенні ранньої діагностики та ефективної профілактики ішемічної хвороби серця.

Робота складається із вступу, шести розділів та висновків.

У першому розділі роботи визначається актуальність даної роботи та описується проблема захворюваності на ішемічну хворобу серця, а також дається короткий опис застосування методів штучного інтелекту та машинного навчання у медичній сфері.

У другому розділі роботи формулюється загальна мета дослідження, а також конкретні завдання, вирішення яких сприятиме досягненню поставленої мети.

У третій частині була описана модель за допомогою багатомірної логістичної регресії.

У четвертому розділі роботи були описані дані, які використовуються для побудови та тестування моделі, а також описується процес обробки цих даних. Також було додано декілька прикладів візуалізації даних.

У п’ятому розділі було розглянуто процес вибору інструментів для реалізації моделі з використанням логістичної регресії.

У шостому розділі була описана реалізація моделі, а також визначені методі оцінки ефективності побудованої моделі.

У результаті виконання роботи була реалізована та проаналізована модель за допомогою багатомірної логістичної регресії, яка передбачає ризик виникнення ішемічної хвороби серця за деякими факторами.

РОЗДІЛ 1. АКТУАЛЬНІСТЬ РОБОТИ

1.1 ІШЕМІЧНА ХВОРОБА СЕРЦЯ

Ішемічна хвороба серця (ІХС) є одним з найпоширеніших та найважливіших захворювань серцево-судинної системи, що виникає внаслідок порушення кровопостачання міокарда через пошкодження та звуження коронарних артерій. Зміни в функціонуванні серцевих судин призводять до порушення балансу між потребами серцевого м'яза та коронарним кровотоком, що призводить до недостатнього окиснювання міокарда та його функціональної недостатності [5].

ІХС частіше діагностується у чоловіків віком від 40 до 45 років, але також може впливати на жінок. Вона може проявлятися у формі стенокардії з періодичними нападами болю в області серця, або у вигляді гострих станів, таких як інфаркт міокарда, що виникає в результаті некрозу м'язової тканини серця. ІХС є однією з провідних причин інвалідизації та смертності серед населення [4], зокрема, близько 40% випадків смертності через патології серця пов'язані з ІХС [1].

Для класифікації ІХС використовуються різні системи [2]. Однією з найбільш відомих є класифікація, запропонована Всесвітньою організацією охорони здоров'я (ВООЗ) у 1984 році [6].

Вона включає такі основні категорії:

1. Раптова коронарна смерть через первинну зупинку серця, яка може бути:

- з успішною реанімацією;
- з летальним кінцем.

2. Стенокардія.

3. Інфаркт міокарда.

4. Постінфарктний кардіосклероз.

5. Порушення серцевого ритму.

6. Серцева недостатність.

Головною причиною розвитку ІХС є атеросклероз коронарних артерій [3]. Це патологічний стан, що поступово розвивається, але швидко прогресує. Атеросклероз може вражати одну або декілька артерій серця, порушуючи їхню просвітність і спричиняючи ішемію міокарда. При значному звуженні просвіту артерії (90-95%) виникає критичне порушення коронарного кровообігу, що може призвести до інфаркту міокарда.

Фактори, що збільшують ризик розвитку ІХС:

- чоловіча стать;
- вік 40-50 років;
- цукровий діабет;
- гіподинамія;
- ожиріння;
- захворювання крові;
- спадковість;
- високий рівень холестерину у крові;
- часті стреси;
- артеріальна гіпертензія;
- куріння;
- зловживання алкоголем та ін.

Незважаючи на значні досягнення в лікуванні та профілактиці, ця хвороба залишається серйозним медичним та соціальним викликом.

За оцінками ВООЗ, щороку в усьому світі через серцеві захворювання помирає 12 мільйонів людей [6]. Половина смертей у Сполучених Штатах Америки, Канаді, Німеччини та інших розвинених країнах пов'язана з серцево-судинними захворюваннями [7]. В Україні ця динаміка є ще гіршою. Ранній прогноз серцево-судинних захворювань може допомогти прийняти рішення щодо зміни способу життя у пацієнтів із високим ризиком і, у свою чергу, зменшити ускладнення від хвороби.

1.2 ШТУЧНИЙ ІНТЕЛЕКТ ТА МАШИННЕ НАВЧАННЯ

Використання методів штучного інтелекту та машинного навчання для визначення ризику виникнення серцевих захворювань є важливим і перспективним напрямком досліджень у медичній сфері. Ці методи дозволяють аналізувати великий обсяг клінічних даних та виявляти складні зв'язки між різними факторами ризику та виникненням серцевих захворювань.

Один з найбільш використовуваних підходів полягає в застосуванні багатомірної логістичної регресії, яка дозволяє прогнозувати ймовірність розвитку серцевих захворювань на основі різних клінічних показників, таких як вік, стать, рівень холестерину, артеріальний тиск тощо. Ця модель дозволяє ідентифікувати фактори, що найбільше впливають на ризик серцевих захворювань та розробляти індивідуалізовані стратегії профілактики та лікування для пацієнтів.

Крім того, машинне навчання дозволяє використовувати різноманітні алгоритми класифікації, такі як дерева рішень, випадкові ліси, метод опорних векторів тощо, для прогнозування ризику виникнення серцевих захворювань. Ці алгоритми можуть виявити складні неявні залежності між вхідними

параметрами та ризиком захворювання, що дозволяє отримати більш точні та надійні прогнози.

Загалом, використання методів штучного інтелекту та машинного навчання в медицині для визначення ризику виникнення серцевих захворювань відкриває нові можливості для попередження та лікування цих небезпечних станів, що сприяє підвищенню ефективності та індивідуалізації медичної допомоги.

Висновок за першою частиною

У даному розділі була описана актуальність використання методів штучного інтелекту та машинного навчання для визначення ризику виникнення хвороб серця, яка визначається кількістю та складністю факторів, які впливають на розвиток серцевих захворювань, а також потребою у зручних та ефективних методах їх прогнозування. Серцево-судинні захворювання є однією з провідних причин смертності у світі, тому розробка ефективних стратегій профілактики та лікування є досить актуальною.

За допомогою методів штучного інтелекту та машинного навчання можна аналізувати великий обсяг клінічних даних та виявляти складні взаємозв'язки між різними факторами ризику та виникненням серцевих захворювань. Це дозволяє розробляти індивідуалізовані та точні моделі для прогнозування ризику захворювання для кожного пацієнта.

Отже, використання методів штучного інтелекту та машинного навчання для визначення ризику виникнення хвороб серця є дуже актуальним і перспективним напрямком досліджень у медичній науці. Впровадження таких інноваційних підходів дозволить покращити діагностику, лікування та профілактику серцевих захворювань, що призведе до зниження їхньої поширеності та покращення якості життя пацієнтів.

РОЗДІЛ 2. МЕТА ТА ЗАДАЧІ РОБОТИ

Мета даної дипломної роботи полягає у розробці моделі за допомогою логістичної регресії, що повинна передбачити ризик виникнення ішемічної хвороби серця за деякими факторами, а також аналізі отриманої моделі.

Задачі, які були сформульовані для досягнення цілей дипломної роботи:

- проаналізувати підходи у сфері штучного інтелекту та машинного навчання для розв’язання задач медичної та клінічної тематики;
- розробити та описати модель за допомогою логістичної регресії;
- обрати дані для реалізації та аналізу розробленої моделі, здійснити обробку даних для застосування методів штучного інтелекту та машинного навчання;
- обрати інструменти для реалізації моделі;
- реалізувати модель;
- проаналізувати ефективність реалізовану модель.

Висновок до другої частини

У даній частині автором було визначено сформульовану мету роботи та виділено задачі для досягнення мети.

РОЗДІЛ 3. РОЗРОБКА МОДЕЛІ

3.1. ЛОГІСТИЧНА РЕГРЕСІЯ

В роботі проводиться розробка моделі за допомогою логістичної регресії, що повинна передбачити ризик виникнення ішемічної хвороби серця за деякими факторами.

Логістична регресія – це статистичний метод, який використовується для прогнозування ймовірності виникнення певної події на основі даних. Вона є ефективним інструментом для розуміння взаємозв'язку між незалежними змінними (факторами, атрибутами) та залежною змінною (подія, що має статистичний характер). Головна ідея логістичної регресії полягає в тому, щоб знайти лінійну функцію, яка може розділити дві або більше категорії даних [20].

З іншого боку, ми будемо розглядати логістичну регресію як алгоритм класифікації машинного навчання, який використовується для прогнозування ймовірності категоріальної залежної змінної від однієї чи декількох незалежних змінних [10].

Моделі машинного навчання – це програми, які можна навчити виконанню складних завдань обробки даних без втручання людини [9].

Моделі машинного навчання, побудовані з використанням логістичної регресії, допомагають організаціям отримувати корисну інформацію зі своїх бізнес-даних. Вони можуть використовувати ці дані для прогнозного аналізу, щоб знизити експлуатаційні витрати, підвищити ефективність та прискорити масштабування [10].

Як працює регресійний аналіз?

Будь-який регресійний аналіз складається з наступних кроків [8]:

1. Визначення питання;
2. Збір даних;
3. Створення та навчання моделі регресійного аналізу;
4. Прогнозування для невідомих значень.

Підходи до логістичного регресійного аналізу

Існує три підходи до логістичного регресійного аналізу, що ґрунтується на результатах залежної змінної [17].

1. Бінарна логістична регресія

Бінарна логістична регресія - це один з типів логістичної регресії, який використовується для прогнозування бінарних або двійкових результатів. У цьому випадку залежна змінна може мати лише два можливих значення, наприклад, «так» або «ні», «позитивний» або «негативний», «вижив» або «не вижив», 1 або 0 [14]. Основна ідея бінарної логістичної регресії полягає в тому, що вона моделює логістичну функцію, яка визначає ймовірність того, що залежна змінна приймає одне з двох можливих значень, в залежності від значень незалежних змінних [20].

Цей вид регресії часто використовується у задачах, де потрібно визначити, чи наступить певна подія або ні, такі як прогнозування ймовірності захворювання на певну хворобу, класифікація електронних листів як «спаму» або «неспаму», прогнозування виживання пасажирів в аварії тощо [15].

2. Багаточленна (багатомірна) логістична регресія

Багатомірна логістична регресія, або поліноміальна логістична регресія, є

розширенням бінарної логістичної регресії та дозволяє прогнозувати результати у випадку, коли існує більше ніж дві категорії. Цей тип регресії широко використовується в класифікаційних задачах, де кількість можливих результатів перевищує два [21].

У багатомірній логістичній регресії для кожної категорії результату створюється окрема логістична модель. Наприклад, якщо є три категорії результату («низький», «середній» та «високий»), то буде створено три моделі для передбачення ймовірності належності до кожної категорії.

Основна ідея полягає в тому, що для кожної категорії результату визначаються ваги незалежних змінних, які показують їх вплив на ймовірність належності до цієї категорії. Потім використовується логістична функція для обчислення ймовірності для кожної категорії, і результат вибирається на основі найвищої отриманої ймовірності.

3. Порядкова логістична регресія

Порядкова логістична регресія застосовується для прогнозування порядкової змінної, що має кілька рівнів або категорій, розташованих у відповідній послідовності. Цей вид регресії використовується у випадках, коли залежна змінна має порядок або ієрархію між своїми рівнями [21].

Методика порядкової логістичної регресії аналогічна звичайній логістичній регресії, проте застосовується у випадках, коли результати можуть мати більше двох можливих рівнів. Наприклад, якщо є категорії «низький», «середній» та «високий» рівень задоволеності клієнтів, то порядкова логістична регресія дозволяє дослідити вплив різних факторів на цей рівень задоволеності.

Для побудови моделі порядкової логістичної регресії використовуються аналогічні методи, що й для звичайної логістичної регресії, з врахуванням порядку категорій. У цьому випадку кількість моделей дорівнює кількості

категорій мінус один, оскільки один рівень виступає як базовий для порівняння.

Застосування порядкової логістичної регресії розповсюджене у різних галузях, таких як маркетингові дослідження, соціологія, медицина тощо. Вона дозволяє врахувати порядок категорій при аналізі та прогнозуванні залежних змінних, що робить її корисною для різних досліджень і прийняття рішень.

3.2 ОПИС МОДЕЛІ

У цій роботі було вирішено будувати модель на основі багатомірної логістичної регресії [16].

Однією з головних причин вибору багатомірної логістичної регресії для створення моделі було те, багатомірна логістична регресія використовується для прогнозування ймовірності належності до однієї з категорій результату, який має більше ніж два можливих рівні, тому за допомогою цієї моделі ми не тільки зможемо визначити, чи має пацієнт ризик виникнення хвороби серця, але й визначити рівень цього ризику.

Математично модель виражається за допомогою логістичної функції:

$$P\left(Y = k \middle| X_1, X_2, \dots, X_p\right) = \frac{e^{(\beta_{0k} + \beta_{1k} * X_1 + \beta_{2k} * X_2 + \cdots + \beta_{pk} * X_p)}}{1 + e^{(\beta_{0k} + \beta_{1k} * X_1 + \beta_{2k} * X_2 + \cdots + \beta_{pk} * X_p)}}.$$

У цій формулі:

Y – - це категоріальна змінна, що відображає результат або категорію ризику (наприклад, рівень ризику виникнення хвороби серця).

k – один з рівнів цієї змінної (наприклад, низький, середній, високий ризик).

p – кількість незалежних змінних, які використовуються для прогнозування результату.

$X_1, X_2, \dots, X_p$ – це незалежні змінні, які використовуються для прогнозування результату.

$\beta_{0k}, \beta_{1k}, \dots, \beta_{pk}$ – це параметри моделі, які оцінюються під час навчання моделі.

Також можна зазначити, що незалежні змінні $X_1, X_2, \dots, X_p$ ми будемо називати атрибутами. Ці атрибути створюють деякий вектор *x*, який складається з цих змінних (атрибутів).

Для нашої моделі кожен атрибут є можливим ризиком для виникнення ішемічної хвороби серця. Дивлячись на ці атрибути, ми можемо визначити, чи є на даний момент у людини ризик виникнення хвороби [13].

Висновок за третьою частиною

У даному розділі була описана логістична регресія як алгоритм класифікації машинного навчання, а також була визначена модель для прогнозування ризику виникнення хвороб серця за допомогою багатомірної логістичної регресії.

РОЗДІЛ 4. ВИБІР ТА ОБРОБКА ДАНИХ ДЛЯ РЕАЛІЗАЦІЇ МОДЕЛІ

4.1 ОПИС ДАНИХ

Для побудови та тестування моделі [12] було обрано CSV файл «To Predict Heart Disease».

Посилання:https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression.

Набір даних загальнодоступний на веб-сайті Kaggle, і він є результатом поточного дослідження серцево-судинної системи жителів міста Фремінгем, штат Массачусетс. Мета класифікації полягає в тому, щоб передбачити, чи є у пацієнта ризик майбутньої ішемічної хвороби серця (ІХС). Набір даних надає інформацію про пацієнтів. Він містить понад 4000 записів і 15 атрибутів.

Опис атрибутів

Кожен атрибут є потенційним фактором ризику. Існують демографічні, поведінкові та медичні фактори ризику.

a) Демографічні:

- Sex – стать пацієнта (Nominal);
- Education – освіта пацієнта (Nominal);
- Age – вік пацієнта (Continuous).

b) Поведінкові:

- Current Smoker – чи є пацієнт курцем на даний момент (Nominal);
- Cigs Per Day – кількість сигарет, які людина викурює в середньому за один день (Continuous).

c) Медичні:

- BP Meds – чи приймав пацієнт ліки від артеріального тиску (Nominal);
- Prevalent Stroke – чи був у пацієнта раніше інсульт (Nominal);
- Prevalent Hyp – чи була у пацієнта гіпертонія (Nominal);
- Diabetes – чи був у пацієнта діабет (Nominal);
- Tot Chol – рівень загального холестерину (Continuous);
- Sys BP – систолічний артеріальний тиск (Continuous);
- Dia BP – діастолічний артеріальний тиск (Continuous);
- BMI – індекс маси тіла (Continuous);
- Heart Rate – частота серцевих скорочень (Continuous);
- Glucose – рівень глюкози (Continuous).

Змінна для прогнозування:

Risk of coronary heart disease CHD – ризик майбутньої ішемічної хвороби серця (Binary: 1 – yes/0 – no).

Для побудови моделі було прийнято рішення виключити атрибут Education з моделі, так як рівень освіти має зовсім незначний вплив на появу ІХС.

4.2. ВІЗУАЛІЗАЦІЯ ДАНИХ

Побудуємо декілька діаграм для візуалізації досліджуваних даних (Код програми на мові програмування Python знаходиться у Додатку 2).

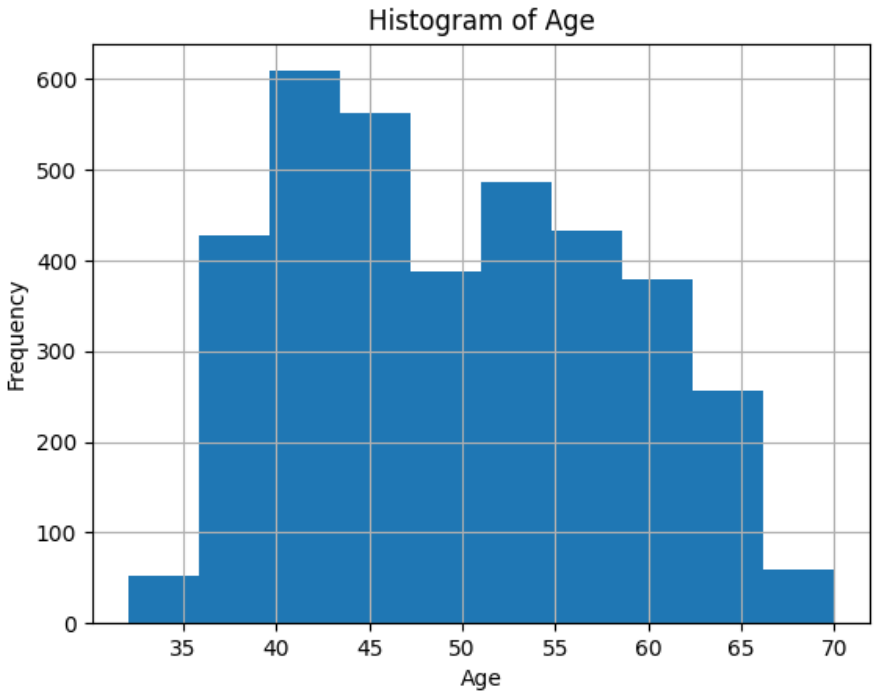

Рисунок 4.2.1. Гістограма розподілу віку

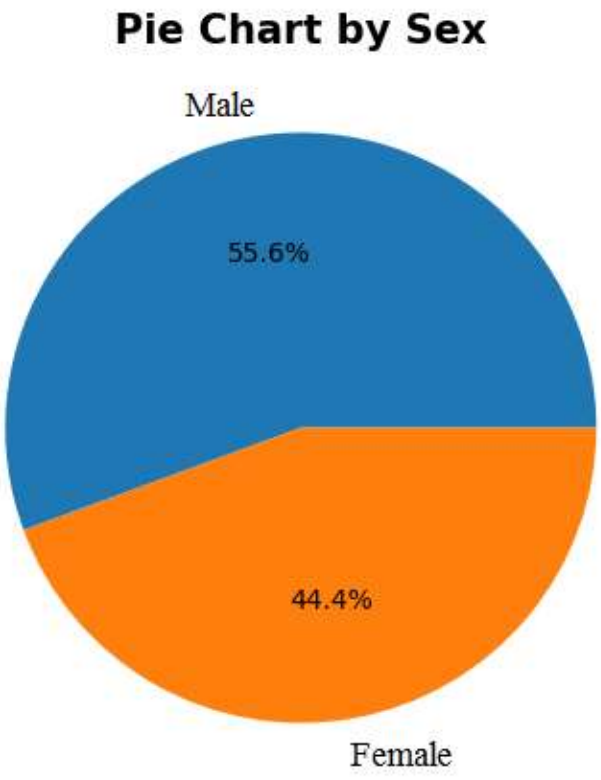

Рисунок 4.2.2. Кругова діаграма розподілу статі

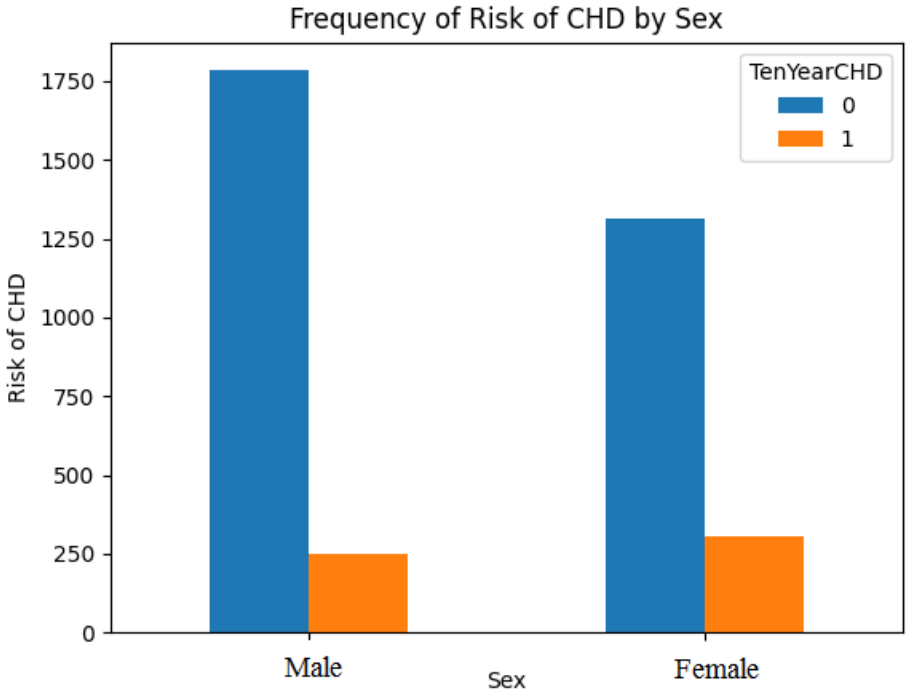

Рисунок 4.2.3. Частота ризику ІХС за статтю

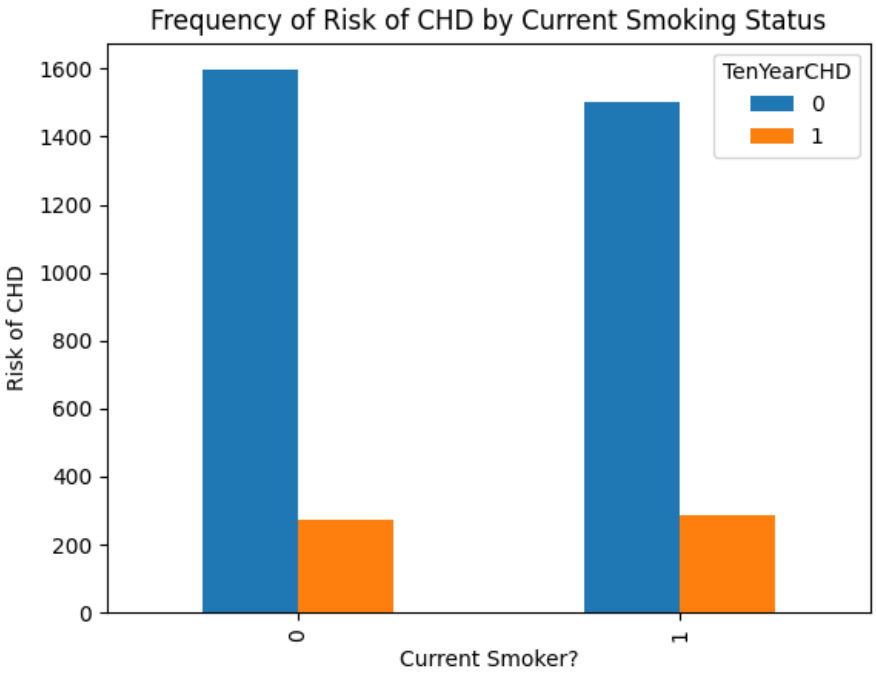

Рисунок 4.2.4. Частота ризику ІХС за статусом курця

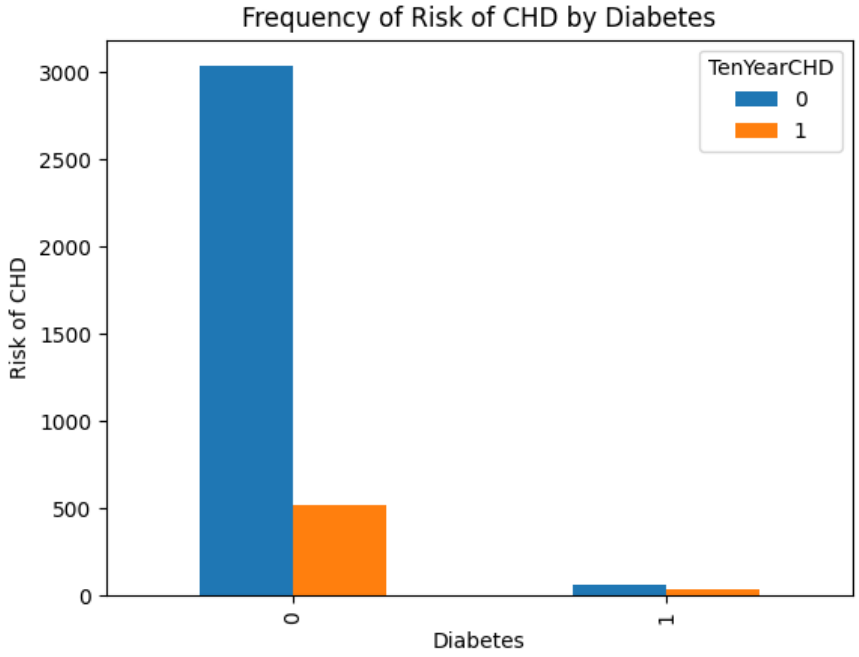

Рисунок 4.2.5. Частота ризику ІХС за фактом того, чи був у пацієнта діабет

4.3 ОБРОБКА ДАНИХ

Ефективна обробка даних є критично важливою передумовою для успішного впровадження методів штучного інтелекту, а також для оптимізації даних перед їх використанням у різноманітних задачах, включаючи застосування логістичної регресії. Цей процес забезпечує створення точних, ефективних та надійних моделей, спроможних вирішувати широкий спектр завдань, що виникають у сфері машинного навчання та аналізу даних.

Нижче наведено кілька кроків, які можна виконати для підготовки даних:

1. **Збір даних**: Збір усіх необхідних даних, які потрібні для моделі. Такі дані можуть включати різні типи змінних, такі як числові, категоріальні та бінарні (Дані для моделі були описані у розділі 4.1.).

2. **Очищення даних**: Перевірка даних на наявність пропущених значень та вирішення цієї проблеми шляхом видалення, заміни або заповнення пропущених значень.

3. **Кодування категоріальних змінних**: Якщо в наборі даних є категоріальні змінні, вони повинні бути перетворені в числові значення. Це може бути зроблено за допомогою методів кодування, таких як one-hot encoding або label encoding [22].

4. **Масштабування числових змінних**: Числові змінні можуть бути масштабовані для полегшення процесу навчання моделі. Зазвичай застосовуються методи масштабування, такі як стандартизація або нормалізація [23].

5. **Розділення на тренувальний та тестовий набори**: Зазвичай вхідні дані розділяються на тренувальний і тестовий набори. Тренувальний набір використовується для навчання моделі, а тестовий – для оцінки її продуктивності.

6. **Балансування класів (за потреби)**: Якщо дані мають дисбаланс класів (коли кількість прикладів в одному класі набагато більша або менша, ніж в іншому), розглядається можливість застосування методів балансування, таких як oversampling або undersampling [24].

7. **Видалення зайвих змінних (за потреби)**: Зазвичай непотрібні або корелюючі змінні, які не приносять користі моделі, видаляються з моделі.

8. **Додавання нових ознак (за потреби)**: За необхідністю є можливість створення нових ознак на основі наявних, які можуть поліпшити результати моделі.

9. **Перевірка кореляції**: Необхідно перевірити кореляцію між змінними та видалити ті, які мають високу кореляцію між собою.

Рисунок 4.3.1. Елементи обробки даних

Зазначені вище кроки допоможуть підготувати дані для побудови моделі, що дозволить отримати кращі результати при її навчанні та оцінці.

Висновок до четвертої частини

Даний розділ описує дані, які будуть використовуватися для реалізації та тестування моделі. Були описані атрибути даних, створена візуалізація деяких критеріїв та атрибутів, а також описані етапи обробки даних для застосування методів штучного інтелекту та машинного навчання.

РОЗДІЛ 5. ВИБІР ІНСТРУМЕНТІВ РЕАЛІЗАЦІЇ МОДЕЛІ

Метою даної роботи є розробка моделі прогнозування ризику виникнення хвороб серця за допомогою багатомірної логістичної регресії. Однак, слід розуміти, що забезпечення стабільності та точності результатів в роботі додатку є надзвичайно важливим завданням. Всі компоненти програми повинні бути розроблені з урахуванням високих стандартів якості. Вибір якісних інструментів розробки визначає успішність проекту і впливає на ефективність та надійність моделі.

Для реалізації моделі прогнозування ризику виникнення ІХС за декількома незалежними змінними була обрана мова програмування Python. Для обробки даних для моделі була застосована бібліотека Pandas, для побудови та аналізу ефективності моделі була застосована бібліотека scilit-learn. Далі ми опишемо кожен з цих компонентів.

Мова програмування Python

Python – це високорівнева мова програмування, яка відома своєю простотою та зручністю використання. Вона використовується для широкого спектру застосувань, включаючи веб-розробку, аналіз даних, штучний інтелект та машинне навчання [22].

Однією з ключових особливостей Python є його читабельний синтаксис, що нагадує природну мову, що робить його ідеальним вибором для початківців у програмуванні та широкого кола фахівців. Python підтримує об'єктно-орієнтований, процедурний та функціональний стилі програмування, що надає розробникам велику гнучкість у створенні програм.

Бібліотека Pandas

Бібліотека Pandas є однією з найпопулярніших інструментів для обробки та аналізу даних в мові програмування Python. Вона надає потужні та зручні інструменти для роботи з даними у вигляді таблиць (або DataFrame), що робить її незамінним інструментом для широкого спектру завдань в області науки про дані, аналізу даних та машинного навчання [19].

Основні функції та можливості бібліотеки Pandas включають:

1. **Створення та завантаження даних**: Pandas дозволяє легко створювати DataFrame з різних джерел даних, таких як CSV-файли, бази даних SQL, Excel-файли, JSON-структури тощо.

2. **Маніпулювання даними**: Завдяки широкому спектру функцій, таких як відбір даних, групування, сортування, злиття та об'єднання таблиць, Pandas дозволяє здійснювати різноманітні операції з даними.

3. **Очищення даних**: Pandas надає зручні інструменти для роботи з пропущеними значеннями та видаленням дублікатів, що допомагає забезпечити якість та консистентність даних.

4. **Візуалізація даних**: Бібліотека має інтеграцію з іншими інструментами візуалізації даних, такими як Matplotlib та Seaborn [19], що дозволяє легко створювати графіки та діаграми для візуального аналізу даних.

5. **Робота з часовими рядами**: Pandas має підтримку для роботи з часовими рядами, включаючи різноманітні функції для аналізу та обробки даних зі специфічними для цього типу даних операціями.

Завдяки своїм потужним можливостям та простоті використання, бібліотека Pandas стала невід'ємною складовою екосистеми Python для роботи з даними та відтворює ключову роль у багатьох проектах в області аналізу даних та машинного навчання.

Бібліотека Scikit-learn (sklearn)

Бібліотека Scikit-learn (sklearn) є однією з найпопулярніших та найважливіших бібліотек для машинного навчання в мові програмування Python. Вона надає простий та ефективний інтерфейс для розробки та навчання різноманітних моделей машинного навчання [18].

Основні функції та можливості бібліотеки Scikit-learn включають:

1. **Відмінну документацію та ресурси**: Scikit-learn має високоякісну документацію, яка включає ясні приклади використання та пояснення параметрів для кожного алгоритму машинного навчання.
2. **Реалізацію різноманітних алгоритмів**: Бібліотека містить реалізації широкого спектру алгоритмів машинного навчання, включаючи класифікацію, регресію, кластеризацію, зменшення розмірності, виявлення аномалій та багато інших.
3. **Простий та консистентний інтерфейс**: Scikit-learn пропонує однаковий інтерфейс для всіх своїх моделей, що спрощує роботу з бібліотекою та дозволяє легко переключатися між різними алгоритмами.
4. **Удосконалені інструменти підготовки даних**: Scikit-learn містить вбудовані інструменти для підготовки даних, включаючи масштабування, кодування категоріальних змінних та видалення аномалій.
5. **Підтримку векторизації та оптимізації для обробки великих обсягів даних**: Бібліотека Scikit-learn надає широкі можливості для векторизації та оптимізації роботи з великими обсягами даних, що дозволяє ефективно працювати над великими наборами даних.

Завдяки своїм потужним інструментам та простому використанню, Scikit-learn є важливою складовою екосистеми Python для розробки моделей машинного навчання та знаходить широке застосування у науці про дані, аналізі даних, промисловості та дослідженнях.

Висновок за п’ятою частиною

У даному розділі було визначено й обґрунтовано, що для розробки додатку будуть використовуватися такі інструменти:

- ✓ Мова програмування Python – реалізація моделі;
- ✓ Бібліотека Pandas – обробка даних для моделі;
- ✓ Бібліотека Scikit-learn – побудова та аналіз ефективності моделі.

РОЗДІЛ 6. РЕАЛІЗАЦІЯ ТА ОЦІНКА МОДЕЛІ

6.1. РЕАЛІЗАЦІЯ МОДЕЛІ

Реалізація моделі багатомірної логістичної регресії – це процес побудови математичної моделі для класифікації даних, коли ми маємо більше ніж два можливі класи або категорії вихідної змінної. Ця модель використовує логістичну функцію для передбачення ймовірності належності спостереження до кожного з класів.

Першим етапом у реалізації моделі є підготовка вхідних даних. Дані зазвичай розділяються на дві частини: тренувальний та тестувальний набори (частини). Тренувальний набір використовується для навчання моделі, тобто для побудови регресійної функції. Він містить вхідні змінні та відповідні значення цільової змінної. Тестувальний набір використовується для оцінки ефективності моделі на незалежних даних. Він також містить вхідні змінні та відповідні значення цільової змінної, але використовується тільки для оцінки точності моделі, а не для її навчання.

Для реалізації розробленої моделі вхідний датасет був розбитий на тренувальну (75%) та тестувальну (25%) частини.

Код програми на мові програмування Python з використанням бібліотек Pandas та Scikit-learn знаходиться у Додатку 1.

6.2 ОЦІНКА МОДЕЛІ

ROC-крива (Receiver Operating Characteristic) [11] – це графічне зображення, яке часто використовується для аналізу результатів бінарної класифікації в машинному навчанні. Назва виникла з систем обробки сигналів.

З огляду на те, що класифікатор розділяє дані на два класи, один з яких має позитивні наслідки, а інший – негативні, ROC-крива відображає залежність між кількістю правильно класифікованих позитивних випадків і кількістю неправильно класифікованих негативних випадків.

У термінології ROC-аналізу перший клас називається істинно позитивною множиною, а другий – хибно негативною. При цьому припускається, що класифікатор має певний параметр, який варіюється, і залежно від нього ми отримуємо різні розділення на два класи. Цей параметр часто називають порогом або точкою відсікання (cut-off value). Від нього залежать різні величини помилок I та II роду. У логістичній регресії поріг відсікання змінюється від 0 до 1 – це розрахункове значення рівняння регресії, яке часто називають рейтингом.

Для розуміння суті помилок I та II роду розглянемо матрицю помилок (confusion matrix), яка будується на основі результатів класифікації моделлю та фактичною (об'єктивною) приналежністю прикладів до класів.

Передбачення моделі	Фактично позитивно	Фактично негативно
Позитивно	TP	FP
Негативно	FN	TN

Таблиця 6.2.1. Матриця помилок (confusion matrix)

TP (True Positives) – це випадки, коли позитивні приклади правильно класифіковані як позитивні (істинно позитивні передбачення). TN (True Negatives) – це випадки, коли негативні приклади правильно класифіковані як негативні (істинно негативні передбачення). FN (False Negatives) – це ситуації, коли позитивні приклади помилково класифікуються як негативні (помилка I роду). Це може статися, наприклад, коли подія, яка насправді має місце,

помилково не виявляється (хибно негативні передбачення). FP (False Positives) – це ситуації, коли негативні приклади помилково класифікуються як позитивні (помилка II роду). Це може виникнути, коли відсутність події помилково визначається як її присутність (хибно позитивні передбачення).

При аналізі частіше оперують не абсолютними показниками, а відносними (rates), вираженими у відсотках:

Частка істинно позитивних передбачень: $TPR = \frac{TP}{TP+FN} * 100\%$.

Частка хибно позитивних передбачень: $FPR = \frac{FP}{TN+FP} * 100\%$.

Введемо ще декілька визначення: чутливість та специфічність моделі. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

Чутливість (Sensitivity) - це і є частка істинно позитивних передбачень:

$$S_e = TPR = \frac{TP}{TP+FN} * 100\%.$$

Специфічність (Specificity) — частка істинно негативних наслідків, які були правильно ідентифіковані моделлю:

$$S_p = \frac{TN}{TN+FP} * 100\%.$$

ROC-крива отримується таким чином:

1) Для кожного значення порога , яке змінюється від 0 до 1 з кроком d_x (наприклад, 0.001) розраховуються значення чутливості S_e та специфічності S_p. В якості альтернативи порогом може бути кожне наступне значення прикладу у виборці.

2) Будується графік залежності: по осі Y відкладається чутливість S_e, по осі X – $FRP = 100 - S_p$ – частка помилково позитивних передбачень.

В результаті вимальовується деяка крива:

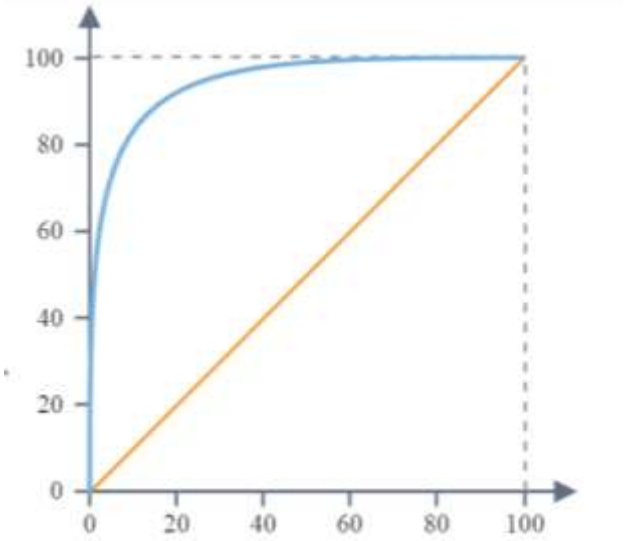

Рисунок 6.2.1. Приклад ROC-кривої

Для ідеального класифікатора ROC-крива перетинає верхній лівий кут, де чутливість (доля істинно позитивних передбачень) досягає 100% або 1,0, а специфічність (доля хибно позитивних передбачень) дорівнює нулю. Отже, чим ближче крива до верхнього лівого кута, тим більше передбачувальна здатність моделі. Навпаки, чим менше вигин кривої та чим ближче вона до діагональної прямої, тим менш ефективна модель. Діагональна лінія відповідає «некорисному» класифікатору, коли відмінностей між двома класами немає.

Один із способів порівняння ROC-кривих – оцінка площі під ними. Теоретично ця площа може змінюватись від 0 до 1, але оскільки модель завжди характеризується кривою, яка розташована вище позитивної діагоналі, то зазвичай говорять про значення від 0,5 («некорисний» класифікатор) до 1 («ідеальна» модель).

Цю оцінку можна отримати безпосередньо обчисленням площі під багатогранником, обмеженим праворуч і знизу осями координат і зліва вгорі. Чисельний показник площі під кривою називається AUC (Area Under Curve).

У літературі іноді наводиться наступна експертна шкала для значень AUC, за якою можна судити про якість моделі [23]:

Інтервал AUC	Якість моделі
0.9 – 1	Відмінна
0.8 – 0.9	Дуже добра
0.7 – 0.8	Добра
0.6 – 0.7	Середня
0.5 – 0.6	Незадовільна

Таблиця 6.2.2. Експертна шкала для значень AUC

Отримана матриця помилок для побудованої моделі:

Передбачення моделі	Фактично позитивно	Фактично негативно
Позитивно	2289	33
Негативно	30	390

Таблиця 6.2.3. Отримана матриця помилок для побудованої моделі

Тобто результат показує, що у нас 2289+390 = 2679 вірних прогнозів та 30+33 = 63 помилкових.

Отримана ROC-крива для побудованої моделі (Код програми на мові програмування Python знаходиться у Додатку 1):

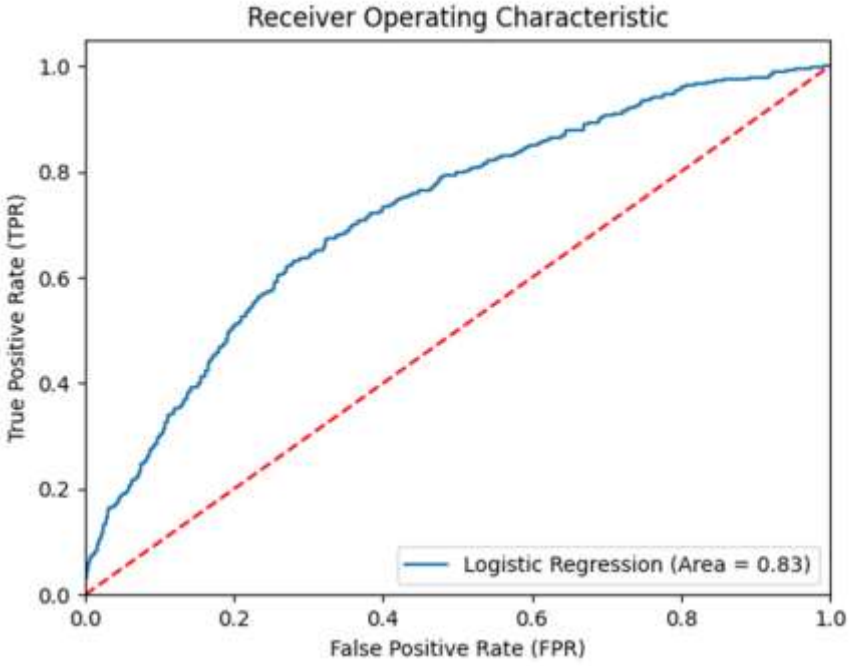

Рисунок 6.2.2. Отримана ROC-крива для побудованої моделі

Також можемо отримати точність моделі, що визначається як відсоток правильних прогнозів, зроблених моделлю:

$$Accuracy = 0.8468271334792122.$$

Це свідчить про те, що модель здійснила вірний прогноз щодо ризику ішемічної хвороби серця у 84,7% випадків.

Висновок до шостої частини

У даному розділі була описана реалізація моделі багатомірної логістичної регресії, а також визначений опис оцінки побудованої моделі.

ВИСНОВКИ

Головною метою даної роботи було розроблення та впровадження моделі на основі логістичної регресії, яка мала передбачати ризик виникнення ішемічної хвороби серця на основі різних факторів, а також проведення аналізу отриманої моделі.

В рамках дипломної роботи було досягнуто наступних результатів.

Були проаналізовані підходи у сфері штучного інтелекту та машинного навчання, які застосовуються в області медицини.

Була розроблена та описана модель з використанням багатомірної логістичної регресії.

Були обрані дані для реалізації та аналізу розробленої моделі, проведена обробка обраних даних для застосування методів штучного інтелекту.

Були обрані та описані інструменти реалізації моделі.

Була реалізована модель, а також був проведений аналіз ефективності реалізованої моделі.

Крім того, були зроблені наступні висновки щодо впливу атрибутів на ризик виникнення ІХС на основі розробленої моделі.

Усі атрибути, вибрані після процесу виключення, показують p-value, нижчі за 5%, що свідчить про значну роль у прогнозуванні захворювань серця.

Чоловіки, здається, більш сприйнятливі до захворювань серця, ніж жінки.

Збільшення віку, кількість викурених сигарет на день і систолічний артеріальний тиск також вказують на збільшення ймовірності захворювання серця.

Загальний рівень холестерину не показує істотних змін у ймовірності ІХС. Це може бути пов’язано з наявністю «хорошого холестерину» у показниках загального холестерину.

Глюкоза також викликає дуже незначну зміну шансів (0,2%)

Модель передбачила з точністю 0,84. Модель більше специфічна, ніж чутлива. Загальну модель можна покращити за допомогою додаткових даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Acute myocardial ischemia in adults secondary to missed Kawasaki disease in childhood / S.R. Rizk, G. El Said, L.B. Daniels [et al.] // Am. J. Cardiol. — 2015. — Vol. 115. — P. 423-427.
2. Серцево-судинні захворювання: Класифікація, схеми діагностики та лікування/ За редакцією професорів В.М. Коваленка, М.І. Лутая. - К. , 2018 . - 77 с.
3. 2018 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). European Heart Journal, Volume 37, Issue 3, 14 January 2018, Pages 267–315
4. Передерій, В. Г. Основи внутрішньої медицини : підруч. для студ. вищ. мед. навч. закл. Т. 2. Захворювання серцево-судинної системи. Загальні питання внутрішньої медицини / В. Г. Передерій, С. М. Ткач. - Вінниця : Нова книга, 2017 – С. 150-224.
5. Slabkiy GO, Parhomenko GY, Astahova NY. Health 2020 - New European Policy and Strategy In the Interest of Health Population. Bulletin of problems in biology and medicine. 2020; Issue 3 (110): 16-20.
6. European action plan for strengthening public health capacities and services. Copenhagen. WHO Regional Office for Europe. 2017. http://www.euro.who.int/_data/assets/pdf_file/0005/171770/ RC62wd12rev1-Eng. pdf).
7. Acheson D. Public health in England: the report of the Committee of Inquiry into the Future Development of the Public Health Function. London, H. M. Stationery Office, 2016
8. Коляда Ю. Фазові та параметричні портрети ключових математичних моделей нелінійної динаміки. [Електронний ресурс]. – Режим доступу: http://www.nbuv.gov.ua/portal/Soc_Gum/Mise/ 2020_82/Kolyda.pdf.
9. Machine Learning: The New AI. - Ethem Alpaydin, 2021.
10. Jason Bronwlee. Logistic Regression for Machine Learning . URL: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

11. Sokolova M. Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Regression Evaluation . Sokolova M., Japkowicz N., Szpakowicz S. – American Association for Artificial Intelligence. - 2020

12. Марценюк В. П. Про програмне середовище проектування інтелектуальних медичних баз даних / В. П. Марценюк, Н. О. Кравець // Клінічна інформатика та телемедицина - 2018. - №> 1. - С. 47-53

13. Кислова О.М., Бондаренко К.Б. Можливості застосування штучних нейронних мереж в аналізі медичної інформації. Вісник Харківського національного університету імені В.Н. Каразіна. 2020. № 891. С. 78-82.

14. Соснін, А. С. Функції активації нейромережі: СІГМОІДА, ЛІНІЙНА, RELU, TAHN. Наука. Інформатизація. Технології. Освіта: матеріали XII міжнародної науково-практичної конференції. Житомир, 2019. С. 237-256.

15. Брандт З. Аналіз даних: Статистичні та обчислювальні методи для науковців та інженерів / З. Брандт. – Київ. АРТ, 2018. – 686 с.

16. Вуколов Е.А. Основи статистичного аналізу. Практикум зі статистичних методів та дослідження операцій з використанням пакетів STATISTICA і EXCEL / Э.А. Вуколов. – Київ. : ФОРУМ, 2019. – 464 с.

17. Вучков И. Прикладний лінійний регресійний аналіз / И. Вучков, Л. Бояджиєва, Е. Солаков. – Харків. : Фінанси і статистика, 2020. – 239 с.

18. Fabian Pedregosa; Gaël Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). "scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*. **12**: 2825–2830.

19. *Molin, Stefanie (2019). Hands-On Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python. Packt. ISBN 978-1-7896-1532-6.*

20. Gareth James, Daniela Witten, Trevor Hastie та Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R.

21. Richard A. Johnson, Dean W. Wichern. Applied Multivariate Statistical Analysis.

22. Sebastian Raschka, Vahid Mirjalili. Python Machine Learning.

23. Andreas C. Müller, Sarah Guido. Introduction to Machine Learning with Python: A Guide for Data Scientists.

24. Haibo He, Yunqian Ma. Imbalanced Learning: Foundations, Algorithms, and Applications.

ДОДАТОК 1. ПОБУДОВА МОДЕЛІ

```
import pandas as pd
import numpy as np
from sklearn.utils import shuffle
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split as tts
from sklearn.metrics import classification_report as CRep
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
import matplotlib.pyplot as plt

'''
1. PREPARATION OF THE DATA
'''

#Read the file with data and drop records with missing value
myData = pd.read_csv("framingham.csv").dropna()

#Define variables
xList = ["male", "age", "currentSmoker", "cigsPerDay", "BPMeds",
         "glucose", "heartRate", "prevalentStroke", "prevalentHyp",
         "diabetes", "totChol", "sysBP", "diaBP", "BMI", "heartRate"]

yVar = "TenYearCHD"

xData = myData[xList]
yData = myData[yVar]

#Split the dataset into training and testing sets (75/25)
xTrain, xTest, yTrain, yTest = tts(xData, yData, test_size = 0.75)


'''
2. CREATE AND TRAIN THE MODEL AND PREDICT THE RESULTS
'''

#Instantiate the model
myModel = LogisticRegression(max_iter = 2000)
#Fit the model using the training data
myModel.fit(xTrain, yTrain)
#Use the model to make predictions on the testing data
yPredict = myModel.predict(xTest)

'''
3. GET THE RESULTS
'''
#Create the confusing matrix
cnfMatrix = metrics.confusion_matrix(yTest, yPredict)
print(cnfMatrix)
#Get an accuracy of the model
print("\n  Accuracy: ", metrics.accuracy_score(yTest, yPredict))

#Get a classification report
print("\n  Classification report: \n\n", CRep(yTest, yPredict))

#Built ROC-curve
logit_roc_auc = roc_auc_score(yTest, yPredict)
fpr, tpr, thresholds = roc_curve(yTest, myModel.predict_proba(xTest)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (Area = %0.2f)' % (logit_roc_auc))
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate (FPR)')
plt.ylabel('True Positive Rate (TPR)')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

ДОДАТОК 2. ПОБУДОВА ГРАФІКІВ

```
import pandas as pd
import matplotlib.pyplot as plt

#Read the file with data and drop records with missing value
myData = pd.read_csv("framingham.csv").dropna()

#Pie Chart by Sex
sex_counts = myData['male'].value_counts()
plt.figure(figsize=(5,5))
labels = sex_counts.keys()
plt.pie(sex_counts.values, labels = labels, autopct = "%1.1f%%")
plt.title("Pie Chart by Sex", fontweight = "bold", fontsize = 15)
plt.savefig('PC_by_Sex')

#Histogram of Age
myData.age.hist()
plt.title('Histogram of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('hist_age')

#Frequency of Risk of CHD by Sex
pd.crosstab(myData.male, myData.TenYearCHD).plot(kind='bar')
plt.title('Frequency of Risk of CHD by Sex')
plt.xlabel('Sex')
plt.ylabel('Risk of CHD')
plt.savefig('risk_CHD_by_Sex')

#Frequency of Risk of CHD by Current Smoking Status
pd.crosstab(myData.currentSmoker, myData.TenYearCHD).plot(kind='bar')
plt.title('Frequency of Risk of CHD by Current Smoking Status')
plt.xlabel('Current Smoker?')
plt.ylabel('Risk of CHD')
plt.savefig('risk_CHD_by_Smoking')

#Frequency of Risk of CHD by Diabetes
pd.crosstab(myData.diabetes, myData.TenYearCHD).plot(kind='bar')
plt.title('Frequency of Risk of CHD by Diabetes')
plt.xlabel('Diabetes')
plt.ylabel('Risk of CHD')
plt.savefig('risk_CHD_by_Diabetes')
```