Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів машинного навчання для прогнозування серцевого нападу
(тема)

Виконав:
здобувач 2 року навчання
групи ІПЗм-23-3
Олексій САНДІН
(Власне ім'я, ПРІЗВИЩЕ)
Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник проф. Олексій ГАЛУЗА
(посада, прізвище)

Допускається до захисту
Зав. кафедри _____ Кирило СМЕЛЯКОВ
(підпис) (Власне ім'я, ПРІЗВИЩЕ)

2025 р.

Харківський національний університет радіоелектроніки
Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Сандіну Олексію Аркадійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи «Дослідження методів машинного навчання для прогнозування серцевого нападу
затверджена наказом по університету від 15.04. 2024р. № 290 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 23.06.2025
3. Вихідні дані до роботи опис досліджуваних методів машинного навчання для передбачення серцевих нападів для проведення досліджень за обраною предметною областю; мови програмування: Python, середовище розробки IntelliJ IDEA 2023.3.3.
4. Перелік питань, що потрібно опрацювати в роботі: аналіз та порівняння існуючих методів машинного навчання, вибір підходящих методів машинного навчання для дослідження, проектування програмної системи для проведення експерементального дослідження, написання програмних рішень, проведення експериментів та аналіз отриманих результатів.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Отримання завдання	16.04.2025	*виконано*
2	Аналіз предметної галузі і постановка задачі	20.04.2025-04.05.2025	*виконано*
3	Аналіз існуючих методів	05.05.2025-10.05.2025	*виконано*
4	Теоретичне дослідження	10.05.2025-20.05.2025	*виконано*
5	Практичне дослідження	21.05.2025-05.06.2025	*виконано*
6	Підготовка пояснювальної записки	05.06.2025-14.06.2025	*виконано*
7	Підготовка презентації та доповіді	15.06.2025	*виконано*
8	Перевірка на плагіат	18.06.2025	*виконано*
9	Нормоконтроль		*виконано*
10	Рецензування		*виконано*
11	Попередній захист		*виконано*
12	Занесення диплома в електронний архів		*виконано*
13	Допуск до захисту у зав. кафедри		*виконано*

Дата видачі завдання 16.04.2025р.

Здобувач _____ Олексій САНДІН
(підпис)

Керівник роботи _____ проф. Олексій ГАЛУЗА
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 74 с., 22 рис., 2 табл., 11 джерел.

МАШИННЕ НАВЧАННЯ, МЕДИЧНІ ДАНІ, ПРОГНОЗУВАННЯ, СЕРЦЕВИЙ НАПАД, ОБРОБКА ДАНИХ, KNN, LOGISTIC REGRESSION, RANDOM FOREST

Об’єктом дослідження є алгоритми машинного навчання, що застосовуються для прогнозування ризику серцевого нападу на основі медичних показників пацієнтів.

Метою роботи є аналіз ефективності застосування різних моделей машинного навчання для раннього виявлення ймовірності серцевого нападу з використанням історичних медичних даних.

Методами дослідження є вивчення сучасних підходів до обробки медичних даних, побудова моделей на основі алгоритмів логістичної регресії, KNN, Random Forest, а також оптимізація моделей шляхом гіперпараметричного налаштування.

У результаті роботи було розроблено програмну систему, яка дозволяє здійснювати тренування моделей а також автоматичний аналіз медичних даних пацієнта, передбачаючи ризик серцевого нападу з високою точністю.

MACHINE LEARNING, PREDICTION, HEART ATTACK, DATA PROCESSING, MEDICAL DATA, LOGISTIC REGRESSION, KNN, RANDOM FOREST

The object of the study is machine learning algorithms used to predict the risk of heart attack based on patients' medical indicators.

The aim of the work is to analyze the effectiveness of using various machine learning models for early detection of the probability of heart attack using historical medical data.

The research methods are the study of modern approaches to medical data processing, building models based on logistic regression algorithms, KNN, Random Forest, and optimizing models through hyperparametric tuning.

As a result of the work, a software system was developed that allows for training models and automatic analysis of patient medical data, predicting the risk of heart attack.

Завідувачу кафедри
ПІ_____
(скорочена назва кафедри)
проф. Кирилу СМЕЛЯКОВУ_
(вчене звання, сласне ім’я, прізвище)

ЗАЯВА
щодо самостійності виконання кваліфікаційної роботи та можливості її публікації (та/або публікації анотації кваліфікаційної роботи) в електронному архіві відкритого доступу ElAr KhNURE

Я, Сандін Олексій Аркадійович, студент гр. ІПЗм-23-3, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів машинного навчання для прогнозування серцевого нападу», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений(на) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

Дата 20.06.2026 Підпис 

ЗМІСТ

ВСТУП

Сучасний етап розвитку медицини характеризується стрімким зростанням обсягів медичних даних та впровадженням інтелектуальних технологій для підтримки прийняття клінічних рішень. Особливої актуальності набуває застосування методів машинного навчання для прогнозування серцево-судинних захворювань, зокрема інфаркту міокарда. Серцевий напад є однією з провідних причин смертності у світі, і раннє виявлення факторів ризику дозволяє зменшити ймовірність критичних наслідків для пацієнта [7]. Використання інтелектуальних алгоритмів аналізу медичних показників відкриває нові можливості для автоматизації діагностики та профілактики таких захворювань.

Проблематика дослідження набуває особливої значущості через потребу у точному та своєчасному прогнозуванні серцевого нападу, що дає змогу медичним працівникам оперативно реагувати на ризики, оптимізувати план лікування та попередити загострення хвороби [6]. Методи машинного навчання, здатні обробляти великі обсяги різнорідних даних - таких як електрокардіограми, аналіз крові, демографічні характеристики - дають змогу створювати гнучкі моделі прогнозування, що адаптуються до індивідуальних особливостей пацієнтів [5].

Основною метою роботи є розробка та аналіз методів прогнозування серцевого нападу з використанням сучасних алгоритмів машинного навчання. Для досягнення цієї мети необхідно виконати такі задачі:

- проаналізувати існуючі моделі прогнозування серцево-судинних захворювань та визначити їхні сильні й слабкі сторони;
- відібрати релевантні медичні ознаки для побудови точних моделей прогнозування;
- розробити та протестувати кілька моделей машинного навчання, зокрема логістичну регресію, KNN і Random Forest;
- оцінити ефективність моделей за допомогою метрик різних метрик.

Об’єктом дослідження є пацієнти з ризиком виникнення серцевого нападу, а також медичні дані, пов’язані з функціонуванням серцево-судинної системи.

Предметом дослідження є методи машинного навчання, які використовуються для побудови моделей прогнозування на основі структурованих медичних даних. Для дослідження застосовуються такі методи:

- алгоритми машинного навчання (логістична регресія,KNN, Random);
- методи попередньої обробки даних та зниження вимірності;
- статистичні методи для аналізу значущості ознак і порівняння ефективності моделей.

Очікується, що у результаті дослідження будуть визначені найбільш ефективні підходи до автоматизованого прогнозування серцевого нападу, що забезпечать високу точність виявлення ризику навіть при обмежених або неповних даних. Застосування цих методів сприятиме персоналізації лікування, зниженню смертності та зменшенню навантаження на медичну систему.

Результати дослідження матимуть практичну цінність для лікарень, кардіологічних клінік та дослідницьких центрів, зацікавлених у використанні інтелектуальних систем для покращення діагностики, профілактики та ведення пацієнтів із серцево-судинними захворюваннями.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз предметної галузі дослідження

Проблема передбачення серцевого нападу є надзвичайно актуальною у сучасній медичній практиці, оскільки інфаркт міокарда входить до переліку найпоширеніших причин смертності у світі. Своєчасне виявлення загрозливого стану пацієнта дозволяє запобігти фатальним наслідкам, знизити вартість лікування та зменшити навантаження на систему охорони здоров’я. У зв’язку з цим виникає необхідність у впровадженні новітніх інформаційних технологій, зокрема методів машинного навчання, які дозволяють автоматизувати процес аналізу медичних даних та приймати рішення на основі статистично обґрунтованих моделей.

Сучасна кардіологія використовує широкий спектр діагностичних методів: електрокардіографію, ехокардіографію, лабораторні дослідження (рівень холестерину, глюкози, тропоніну тощо), а також оцінку факторів способу життя (куріння, фізична активність, ІМТ тощо). Однак, навіть за наявності великої кількості даних, виявлення закономірностей та прогнозування на основі традиційного аналізу часто є недостатньо ефективним через складність медичних взаємозв’язків і багатофакторну природу серцевих захворювань. Саме тому машинне навчання виступає потужним інструментом, здатним виявити приховані залежності та надати високу точність прогнозування.

Існує декілька категорій підходів, які використовуються для передбачення серцевих нападів. Одним із найбільш точних методів є застосування клінічних шкал ризику, таких як TIMI чи GRACE, які базуються на медичних показниках. Хоча ці методи перевірені на практиці, вони мають обмежену гнучкість і не враховують складні нелінійні взаємозв’язки між ознаками. Крім того, застосування таких шкал вимагає ретельного заповнення параметрів вручну, що обмежує їхню ефективність в умовах великого потоку пацієнтів.

Іншою категорією підходів є статистичні моделі та регресійний аналіз, які дозволяють оцінювати ймовірність серцевого нападу на основі окремих факторів

ризику. Такі моделі добре інтерпретуються лікарями, але поступаються в точності більш складним алгоритмам, які можуть обробляти великі обсяги даних.

Сучасні дослідження зосереджені на використанні алгоритмів машинного навчання – таких як логістична регресія, Random Forest[1], градієнтний бустинг (XGBoost [2], LightGBM[3]), а також згорткові (CNN) та рекурентні (LSTM) нейронні мережі. Ці методи здатні аналізувати як структуровані табличні дані, так і часові ряди (наприклад, ЕКГ-сигнали), демонструючи високу точність при правильному підборі ознак та налаштуванні гіперпараметрів. Вони забезпечують можливість обробки великих масивів історичних даних, побудови адаптивних моделей та виявлення складних залежностей, які не очевидні для людського аналізу.

Однією з переваг впровадження методів машинного навчання у кардіологічну практику є можливість розгортання моделей у вигляді програмних систем, доступних як у локальній інфраструктурі, так і в хмарному середовищі[4]. Завдяки цьому можна досягти швидкої інтеграції з існуючими медичними інформаційними системами, спростити процес прийняття рішень для лікарів, а також забезпечити персоналізований підхід до профілактики та лікування.

Таким чином, аналіз предметної галузі показує, що впровадження машинного навчання у сферу прогнозування серцевих нападів має вагоме практичне значення. Воно дозволяє підвищити точність діагностики, скоротити час на обробку медичної інформації та підвищити ефективність профілактичних заходів, спрямованих на збереження життя пацієнтів.

1.2 Огляд існуючих підходів та їх ефективності

У сфері прогнозування серцевих нападів існує чимало прикладів застосування програмних систем та методів машинного навчання, які дозволяють здійснювати автоматичну обробку медичних даних і на основі цього робити висновки щодо ризику виникнення інфаркту міокарда. Однією з найвідоміших платформ є IBM Watson Health, яка інтегрує аналітичні інструменти для обробки електронних медичних записів, лабораторних результатів і діагностичних даних.

Watson використовує алгоритми машинного навчання, зокрема логістичну регресію та дерева рішень, для оцінки ймовірності розвитку серцево-судинних захворювань. Однак, система має обмеження, пов’язані з високою вартістю впровадження, а також залежністю від великого обсягу якісно структурованих даних.

Інша популярна система – CardioRisk, розроблена як хмарне рішення для аналізу факторів ризику серцевих нападів. Вона застосовує алгоритми класифікації на основі навчання з учителем, включаючи метод опорних векторів та Random Forest. CardioRisk дозволяє проводити швидкий аналіз історії хвороби, віку, рівня холестерину та інших параметрів. Але ця система має обмеження щодо інтерпретованості моделей, що може ускладнити її використання лікарями без спеціальної підготовки в області аналітики.

Також слід згадати приклад Myocardial Infarction Predictive System (MIPS) – системи, що була розроблена в рамках дослідницького проєкту, і базується на градієнтному бустингу. Використовуються великі набори ознак, включаючи демографічні, клінічні та поведінкові фактори, що дозволяє підвищити точність прогнозування. Одним із недоліків таких моделей є «чорний ящик» у прийнятті рішень, коли навіть високоточна модель не може надати зрозуміле пояснення свого результату.

Таблиця 1.1 – Недоліки програмних систем (таблиця виконана самостійно)

Назва системи	Недоліки
IBM Watson Health	- висока вартість ліцензування та впровадження; - потреба в чистих і повних даних;
CardioRisk	- складність інтерпретації результатів для лікарів без технічної освіти; - обмеження у кастомізації моделей;
MIPS	- відсутність прозорості в ухваленні рішень; - складна побудова моделі та значні обчислювальні витрати.

Хоча сьогодні існує широкий спектр інструментів для прогнозування серцевих нападів, усі вони мають певні обмеження – як технічного, так і практичного характеру. Це створює простір для подальших досліджень, зокрема у напрямку побудови моделей, що поєднують високу точність з прозорістю, простотою інтеграції та адаптацією до різних клінічних умов.

1.3 Перспективи та актуальність дослідження

На тлі глобального зростання кількості захворювань серцево-судинної системи, що залишаються однією з головних причин смертності у світі, дослідження методів прогнозування серцевих нападів набуває особливої актуальності. Потреба в надійних, точних та автоматизованих системах, здатних виявити загрозу інфаркту на ранніх стадіях, зумовила активне застосування методів машинного навчання (ML) у кардіологічній практиці. Такі технології відкривають нові можливості у персоналізованій медицині, покращенні діагностики, а також у зниженні навантаження на медичні системи завдяки ранньому втручанню.

Однією з ключових перспективних тенденцій є перехід від класичних статистичних підходів до адаптивних моделей машинного навчання, які здатні враховувати складні, неочевидні взаємозв’язки між клінічними показниками пацієнтів. Алгоритми, такі як логістична регресія, випадковий ліс (Random Forest), метод k-найближчих сусідів (KNN), а також нейронні мережі, забезпечують різні підходи до обробки медичних даних, даючи змогу адаптувати моделі до конкретних груп пацієнтів. Розробка гібридних систем, які поєднують переваги кількох алгоритмів, розглядається як ефективний напрямок покращення точності передбачення.

Ще однією важливою тенденцією є інтеграція багатомодальних джерел даних. Окрім стандартних медичних параметрів, таких як вік, тиск, рівень холестерину чи ЕКГ, все частіше у дослідження включаються такі дані, як аналіз зображень серця, результати лабораторних тестів, генетична інформація або поведінкові патерни. Обробка таких різнорідних даних вимагає використання

сучасних ML-підходів, включаючи ансамблеві моделі, а також методи зменшення розмірності та автоматичного виділення релевантних ознак (feature engineering).

Значну роль у майбутньому матиме розвиток моделей, здатних працювати в режимі реального часу, зокрема на базі переносних пристроїв та систем безперервного моніторингу. Такі рішення, які інтегруються з фітнес-браслетами, смарт-годинниками або вбудованими сенсорами, можуть надавати попереджувальні сигнали при виявленні аномалій у серцевому ритмі, що потенційно вказують на ризик інфаркту. Для цього необхідні легковагові моделі з високою точністю, здатні працювати в умовах обмежених обчислювальних ресурсів – наприклад, із застосуванням моделей градієнтного бустингу або згорткових нейронних мереж із оптимізованою архітектурою.

Особливу увагу дослідники приділяють персоналізації прогнозу. Системи, що враховують унікальні фізіологічні та соціальні фактори кожного пацієнта, здатні надавати індивідуалізовані оцінки ризику. Машинне навчання дозволяє динамічно адаптувати моделі до нових даних пацієнта, враховуючи зміни в стилі життя, медикаментозному лікуванні чи перебігу хвороб. Самонавчальні системи, які удосконалюються в процесі експлуатації завдяки зворотному зв’язку з лікарем або даними з носимих пристроїв, відкривають нові горизонти в довгостроковому моніторингу здоров’я.

Практична значущість впровадження методів машинного навчання в кардіологію полягає не лише в можливості раннього виявлення серцевих нападів, а й у запобіганні ускладненням, зниженні вартості лікування та підвищенні якості життя пацієнтів. Це особливо актуально для пацієнтів із високими факторами ризику, яким потрібне постійне спостереження. Машинне навчання дозволяє автоматизувати аналіз великих масивів даних, що в ручному режимі було б практично неможливо, забезпечуючи при цьому об'єктивність та незалежність від людського чинника.

Таким чином, дослідження методів машинного навчання для передбачення серцевого нападу є не лише актуальним, а й критично важливим напрямком у розвитку сучасної медицини. Його перспективи пов’язані із впровадженням

інтелектуальних систем підтримки рішень, створенням індивідуалізованих програм моніторингу та запобігання, а також з інтеграцією технологій у повсякденну клінічну практику. У майбутньому подібні системи можуть стати основою персоналізованої превентивної медицини, що дозволить зберегти життя тисячам людей.

1.4 Огляд основних джерел

Дослідження методів машинного навчання для прогнозування серцевого нападу показує, що більшість сучасних підходів базується на застосуванні алгоритмів глибокого навчання, ансамблевого моделювання та методів попередньої обробки даних. Основна увага приділяється інтеграції різних моделей, аналізу важливих характеристик та використанню практичних інструментів для забезпечення надійності та точності прогнозів.

У роботі "Heart Attack Prediction Using Machine Learning Techniques: A Review" (2021) [5] розглянуто різноманітні алгоритми машинного навчання, такі як Logistic Regression, Random Forest, Gradient Boosting і K-Nearest Neighbors. Ключовим внеском цього дослідження є систематизація підходів до виявлення найважливіших характеристик пацієнта: віку, індексу маси тіла, рівня холестерину, артеріального тиску, наявності шкідливих звичок та історії захворювань. Особливу увагу приділено оптимізації моделей шляхом використання метрик, таких як точність, чутливість і специфічність, для об’єктивної оцінки їх ефективності.

У статті "Development of Heart Attack Prediction Model Based on Machine Learning" (2023) [6] представлено результати використання методу Support Vector Machine (SVM) для класифікації ризику серцевого нападу. Особливістю дослідження є впровадження технік зменшення дисбалансу даних, таких як синтетична генерація менш представлених зразків за допомогою SMOTE (Synthetic Minority Over-sampling Technique)[9]. Це забезпечило досягнення точності моделі в межах 92%, що свідчить про високу надійність такого підходу.

Важливо, що робота демонструє необхідність глибокого аналізу якості вхідних даних і їх впливу на остаточний прогноз.

Дослідження "Explainable Ensemble Learning Models for Early Detection of Heart Disease" (2024) [7] зосереджується на розробці ансамблевих моделей, які інтегрують рішення Decision Trees і Neural Networks. Унікальність цього підходу полягає у підвищенні стійкості моделі до помилкових класифікацій через поєднання переваг кожного з методів. Експериментальні результати свідчать про значне підвищення метрики AUC-ROC та зменшення кількості хибно-позитивних результатів, що є важливим для клінічного застосування. Також виявлено, що такі моделі мають більшу адаптивність до гетерогенних даних пацієнтів, що дозволяє використовувати їх для міжрегіональних порівнянь.

Робота "ECG Based Early Heart Attack Prediction Using Neural Networks" (2022) [8] демонструє переваги штучних нейронних мереж (ANN) при аналізі великих медичних даних. Автори зосередилися на використанні алгоритмів градієнтного спуску та backpropagation для навчання моделі. Крім того, дослідження акцентує увагу на проблемах перенавчання і використовує методи регуляризації, такі як L2-норма, для покращення узагальнювальних властивостей моделі. Показано, що такі мережі можуть досягати високих показників точності (близько 96%), одночасно знижуючи кількість невизначених випадків, які можуть стати проблемою в медицині.

Всі розглянуті джерела одностайно підкреслюють важливість попередньої обробки даних, що включає:

- очищення даних від пропущених або некоректних значень;
- нормалізацію показників для забезпечення відповідного розподілу;
- вилучення найважливіших характеристик пацієнтів для зменшення шуму в даних і підвищення ефективності моделей.

Важливо також зазначити, що використання трансферного навчання може суттєво поліпшити результати аналізу, оскільки цей підхід дозволяє адаптувати моделі до нових наборів даних без необхідності додаткового навчання на великих зразках.

Таким чином, розглянуті в літературі підходи демонструють високий потенціал для практичного застосування в системах раннього виявлення серцевого нападу. Зокрема, поєднання ансамблевого навчання, нейронних мереж і технік попередньої обробки даних дозволяє підвищити точність та адаптивність моделей, знижуючи ризики і забезпечуючи своєчасне реагування на потенційно небезпечні стани пацієнта.

1.5 Існуючі обмеження

Існуючі рішення у сфері прогнозування серцевих нападів із використанням методів машинного навчання демонструють перспективні результати, однак вони не є досконалими та мають ряд обмежень. Одним з ключових чинників є якість і повнота вхідних медичних даних. Часто моделі навчаються на обмежених або незбалансованих наборах даних, де певні групи пацієнтів (наприклад, за віком, статтю або наявністю хронічних захворювань) представлені недостатньо. Це призводить до упередженості моделей і погіршення їхньої здатності узагальнювати результати на нові випадки, особливо в умовах клінічної практики.

Ще одним важливим обмеженням є різнорідність джерел даних. Пацієнтські медичні записи часто містять пропущені значення, нечіткі або несумісні формати, що ускладнює попередню обробку і знижує якість моделі. У деяких випадках моделі можуть некоректно інтерпретувати дані, отримані в інших клініках або країнах, через відмінності в медичних протоколах, обладнанні або стилі документування.

Обмеженням є також обмежена інтерпретованість складних моделей. Наприклад, ансамблеві методи або глибокі нейронні мережі, хоча й забезпечують високу точність, часто функціонують як “чорні скриньки”, що ускладнює пояснення рішень лікарю або пацієнту. В умовах медицини, де прозорість та довіра до діагнозу мають критичне значення, така непрозорість може знизити готовність до використання моделей у клінічній практиці.

Ще одним суттєвим викликом є персоналізація моделей. Більшість алгоритмів створені на основі усереднених характеристик великої кількості

пацієнтів, але не завжди враховують індивідуальні особливості – генетичні фактори, спосіб життя, або історію хвороб. Це може призводити до того, що передбачення в окремих випадках будуть малоефективними або навіть хибними.

Нарешті, з практичної точки зору, інтеграція моделей машинного навчання в медичні системи стикається з такими труднощами, як технічна сумісність із наявними електронними медичними картками, необхідність дотримання стандартів захисту персональних даних (наприклад, GDPR або HIPAA), а також забезпечення надійної роботи моделей у режимі реального часу. Крім того, у деяких випадках виникають етичні питання, пов’язані з автоматизованим прийняттям рішень без участі лікаря, особливо у випадках високого ризику для життя.

Таким чином, попри очевидні переваги, існуючі методи машинного навчання виявляють низку обмежень, які потребують подальших досліджень, оптимізації та адаптації до умов реальної медичної практики.

1.6 Постановка задачі

Метою цього курсового проєкту є дослідження та розробка ефективного методу прогнозування ризику серцевого нападу з використанням сучасних алгоритмів машинного навчання. Своєчасне передбачення потенційної загрози інфаркту міокарда є критично важливим для зниження рівня смертності та покращення якості життя пацієнтів. Впровадження моделей машинного навчання в клінічну практику дозволяє автоматизувати виявлення високого ризику та оперативно реагувати на зміни у стані здоров’я пацієнта, забезпечуючи профілактичні заходи ще до виникнення гострої ситуації.

Попередні дослідження у цій галузі показали, що різні методи машинного навчання, включно з логістичною регресією, деревами рішень, методами ансамблю та нейронними мережами, можуть забезпечувати високу точність передбачення, проте багато з існуючих рішень демонструють обмежену інтерпретованість, недостатню персоналізацію та слабку узагальнюваність при застосуванні до нових або неоднорідних медичних даних. У зв’язку з цим виникає

потреба не лише у використанні існуючих підходів, а й у розробці гнучкої та ефективної моделі, яка буде враховувати особливості клінічних сценаріїв та різноманіття пацієнтських профілів.

Для досягнення цієї мети необхідно вирішити низку задач. По-перше, провести детальний аналіз існуючих методів прогнозування серцевих нападів із використанням машинного навчання. Необхідно розглянути переваги й недоліки кожного підходу, оцінити можливість їхньої адаптації до клінічних даних, а також визначити найбільш перспективні моделі з точки зору точності, чутливості та інтерпретованості.

Наступним етапом є побудова моделі прогнозування. Потрібно сформувати якісний набір даних з обраними клінічними параметрами (такими як вік, рівень холестерину, артеріальний тиск, рівень глюкози, дані ЕКГ тощо) і застосувати до нього різні алгоритми машинного навчання, зокрема логістичну регресію, випадковий ліс та KNN. Важливо забезпечити оптимізацію гіперпараметрів та збалансування даних, особливо у випадку наявності класової нерівноваги.

Далі необхідно провести інтеграцію та оцінювання моделі. Після навчання моделі слід протестувати її на незалежному наборі даних, визначити метрики якості, такі як точність, повнота, F1-міра, ROC-AUC, і порівняти отримані результати з базовими методами. Особливу увагу потрібно приділити здатності моделі виявляти випадки високого ризику з мінімальною кількістю хибнонегативних результатів.

Наступним кроком є верифікація та удосконалення результатів. На цьому етапі здійснюється аналіз помилкових прогнозів моделі, внесення коректив у підхід до обробки даних.

Завершальним етапом є оцінка практичної ефективності запропонованого підходу. Потрібно визначити потенціал інтеграції моделі в існуючі медичні системи, оцінити її корисність для клініцистів у процесі ухвалення рішень, а також вивчити перспективи масштабування системи для використання в різних медичних установах.

2 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ

2.1 Вибір датасету

Якість і структура навчального набору даних є одним із ключових факторів успішності моделей машинного навчання, особливо в задачах, що стосуються медичної діагностики та прогнозування, таких як передбачення серцевого нападу. Медичні дані часто характеризуються складністю, неоднорідністю, наявністю пропущених значень і потенційним дисбалансом між класами. Тому вибір відповідного датасету має критичне значення для побудови надійної та клінічно значущої моделі. У цьому дослідженні для навчання та оцінювання моделей обрано Heart Disease Classification Dataset, який є відкритим набором даних, доступним на платформі Kaggle, та широко використовується в академічних і прикладних дослідженнях.

Heart Disease Classification Dataset містить структуровані дані пацієнтів з різними медичними показниками, що мають значення для оцінки ризику серцевого нападу. Кожен запис у цьому наборі даних представляє інформацію про окремого пацієнта, включаючи як демографічні характеристики (вік, стать), так і клінічні параметри (артеріальний тиск, рівень холестерину, частота серцевих скорочень, рівень цукру в крові натще, наявність болю в грудях, електрокардіографічні дані тощо). Цільова змінна вказує на наявність або відсутність ризику серцевого нападу, що дозволяє формалізувати задачу як задачу бінарної класифікації.

Даному датасету притаманна важлива властивість – він відносно збалансований, тобто має приблизно однакову кількість прикладів для обох класів (наявність/відсутність хвороби). Це позитивно впливає на стабільність навчання моделі та дозволяє уникнути серйозного перекосу в бік одного з класів, що є типовою проблемою для медичних наборів даних. Однак, незважаючи на це, попередня обробка залишається необхідною, оскільки набір може містити пропущені або аномальні значення, які потребують очищення.

При виборі Heart Disease Classification Dataset враховувалися кілька ключових факторів. Насамперед, набір містить релевантні характеристики, що

описують серцево-судинний стан пацієнтів, і дозволяє будувати моделі з високим потенціалом для клінічного застосування. По-друге, його доступність і структура сприяють швидкому впровадженню моделей та повторюваності експериментів. Нарешті, цей набір вже використовувався у низці досліджень, що дозволяє порівняти результати з іншими підходами та алгоритмами машинного навчання, оцінити доцільність запропонованих моделей і виконати якісну валідацію.

Підготовка даних для подачі в модель машинного навчання включає кілька етапів. Спочатку проводиться очищення даних – виявлення та видалення або заповнення пропущених значень, а також видалення очевидних викидів. Далі виконується кодування категоріальних змінних (наприклад, за допомогою one-hot encoding або label encoding), оскільки багато алгоритмів машинного навчання не працюють безпосередньо з нечисловими ознаками. Числові змінні можуть бути нормалізовані або стандартизовані для покращення збіжності моделей, особливо якщо застосовуються алгоритми, чутливі до масштабування, як-от логістична регресія або метод опорних векторів.

Для збільшення точності та стійкості моделі також передбачено застосування методів зменшення розмірності, таких як аналіз головних компонент (PCA), що дозволяє зменшити надлишковість ознак, а також методів боротьби з дисбалансом, зокрема, через синтетичне додавання прикладів класу меншості (наприклад, SMOTE). Усі ці етапи підготовки спрямовані на забезпечення оптимальної якості вхідних даних, що є критичним чинником для отримання точних та надійних прогнозів у медичній сфері.

Таким чином, вибір Heart Disease Classification Dataset як основного джерела даних у цьому дослідженні обумовлений його відповідністю поставленій задачі, структурною повнотою, наявністю релевантних ознак і широким визнанням у спільноті. Це створює надійну основу для побудови, навчання та тестування моделей машинного навчання, здатних ефективно прогнозувати ризик серцевого нападу на основі клінічних даних пацієнта.

2.2 Огляд існуючих методів

У сучасній медичній інформатиці значна увага приділяється розробці методів прогнозування серцевого нападу з використанням алгоритмів машинного навчання (ML). Це зумовлено потребою в ранньому виявленні ризиків серцево-судинних захворювань, що дозволяє своєчасно вжити профілактичних заходів та зменшити смертність. Застосування ML-методів відкриває можливості для побудови моделей, які здатні аналізувати великі обсяги клінічних даних та виявляти складні взаємозв'язки між ознаками, що важко помітити за допомогою традиційної статистики.

До найбільш поширених методів машинного навчання, що застосовуються для передбачення серцевого нападу, належать логістична регресія (Logistic Regression), метод найближчих сусідів (K-Nearest Neighbors, KNN), випадковий ліс (Random Forest), градієнтний бустинг (Gradient Boosting) та нейронні мережі (Artificial Neural Networks). Кожен з цих методів має свої особливості, переваги та недоліки, які визначають доцільність їхнього використання в конкретних умовах.

Логістична регресія є класичним статистичним методом, адаптованим для бінарної класифікації. Її основною перевагою є інтерпретованість та швидкість навчання. Вона добре працює в задачах, де передбачуваний результат – це ймовірність настання події (наприклад, інфаркту) за наявності певного набору факторів ризику. Проте, логістична регресія має обмежену здатність до моделювання нелінійних залежностей і може поступатися більш складним моделям при обробці великої кількості ознак або взаємозв'язків між ними.

Метод найближчих сусідів (KNN) працює на основі принципу подібності: прогноз для нового прикладу здійснюється на основі результатів найближчих за ознаками пацієнтів у навчальній вибірці. Попри свою простоту та інтуїтивність, KNN є чутливим до вибору метрики відстані, розміру вибірки та масштабу ознак. Його продуктивність швидко погіршується при високій розмірності простору ознак (так зване “прокляття розмірності”), а також за наявності шуму в даних.

Випадковий ліс (Random Forest) є ансамблевим методом, який поєднує велику кількість дерев рішень, кожне з яких навчається на випадковій підмножині

ознак та прикладів. Це забезпечує високу стійкість до перенавчання та дозволяє моделі враховувати складні, нелінійні взаємозв'язки між параметрами пацієнта (вік, артеріальний тиск, рівень холестерину тощо). Random Forest також має вбудовані механізми оцінки важливості ознак, що дозволяє краще розуміти, які параметри найбільше впливають на ймовірність серцевого нападу. Недоліком цього методу є складність інтерпретації результатів та відносно високі обчислювальні витрати при роботі з великими наборами даних.

Градієнтний бустинг (наприклад, XGBoost, LightGBM) продовжує розвиток ідей ансамблевих методів, створюючи сильну модель шляхом послідовного додавання слабких моделей (зазвичай дерев рішень), кожна з яких намагається компенсувати помилки попередніх. Цей метод показав виняткову ефективність у багатьох медичних задачах, зокрема у класифікації серцево-судинних ризиків. Проте, його налаштування вимагає ретельного добору гіперпараметрів та обережності щодо перенавчання, особливо на невеликих вибірках.

Штучні нейронні мережі (Artificial Neural Networks, ANN) дозволяють моделювати складні та багаторівневі залежності між ознаками. У порівнянні з класичними ML-методами, нейронні мережі мають здатність автоматично вилучати приховані закономірності, що робить їх ефективними для задач зі складною структурою даних. Проте їхня головна проблема – це необхідність у великому обсязі навчальних даних, висока обчислювальна складність, а також обмежена інтерпретованість моделі, що може бути критичним у сфері охорони здоров’я.

Особливу роль у сучасних підходах відіграє перенесення навчання та автоматичний підбір параметрів (наприклад, з використанням GridSearchCV чи RandomizedSearchCV), що дозволяють значно підвищити точність моделей навіть при обмеженій кількості даних. Також дедалі частіше використовуються гібридні підходи, які поєднують декілька методів з метою досягнення більшої точності та стійкості моделей.

Порівнюючи методи машинного навчання з традиційними медичними підходами до оцінки ризику серцевих нападів (наприклад, шкала Фремінгема),

слід відзначити, що ML-моделі здатні адаптуватися до нових типів даних, обробляти складні мультифакторні залежності та працювати в режимі реального часу. Вони не тільки автоматизують процес діагностики, але й забезпечують індивідуалізований підхід до пацієнта. Проте їхнє використання потребує ретельної валідації, дотримання етичних норм і стандартів безпеки даних.

У межах даної роботи планується застосування декількох класичних методів машинного навчання, таких як логістична регресія, Random Forest та KNN. Кожен з методів буде порівнюватися за точністю, повнотою та іншими метриками продуктивності на основі відкритого медичного датасету. Особлива увага приділятиметься підбору оптимальних параметрів моделей, аналізу важливості ознак та перевірці стійкості моделей до змін у даних.

2.3 Порівняння моделей

Після аналізу найбільш популярних підходів машинного навчання, що застосовуються для медичної діагностики, доцільно провести порівняльний аналіз моделей, обраних для дослідження. Це дозволить визначити їхні переваги та обмеження у контексті завдання прогнозування серцевого нападу. У рамках даної роботи було обрано такі моделі: логістична регресія (Logistic Regression), метод k-ближчих сусідів (KNN) та випадковий ліс (Random Forest). Їх вибір обґрунтовується поєднанням інтерпретованості, ефективності при обмежених даних і можливістю масштабування для складніших задач. Для систематизації ключових характеристик моделей подано порівняльну таблицю.

Кожна з цих моделей має свої сильні сторони та обмеження. Логістична регресія вирізняється високою інтерпретованістю – її коефіцієнти дають змогу оцінити вплив кожної ознаки на ймовірність настання серцевого нападу. Це особливо важливо у медичних задачах, де лікарі повинні розуміти логіку прийняття рішення. Проте лінійна природа цієї моделі може бути недостатньою для складних нелінійних залежностей у даних.

Таблиця 2.3.1 – Порівняльний аналіз моделей машинного навчання (таблиця виконана самостійно)

Параметр	Логістична регресія	K-ближчих сусідів (KNN)	Випадковий ліс (Random Forest)
Основна ідея	Лінійна модель для класифікації ймовірності події	Класифікація на основі близькості до навчальних зразків	Ансамбль рішень дерев для зменшення переобучення
Складність навчання	Низька	Відсутня фаза навчання (високі витрати при прогнозі)	Середня до високої (багато дерев)
Чутливість до масштабування	Висока	Висока (необхідна нормалізація)	Низька (внутрішня стійкість)
Інтерпретованість	Висока	Низька	Середня
Схильність до перенавчання	Низька	Висока при малих значеннях k	Низька завдяки ансамблюванню
Час прогнозування	Дуже швидкий	Повільний (залежить від обсягу даних)	Швидкий
Стійкість до шумів	Низька	Низька	Висока
Придатність до медичних задач	Висока (простота й прозорість)	Середня (потребує налаштування)	Висока (точність і гнучкість)

Метод k-ближчих сусідів є інтуїтивно зрозумілим і простим у реалізації. Він не вимагає навчання у традиційному сенсі, натомість приймає рішення на основі схожості нових прикладів до наявних. Проте при великих обсягах даних час

класифікації значно зростає, що обмежує його практичну застосовність у реальному часі. Також модель чутлива до шумів та до вибору гіперпараметра k.

Натомість випадковий ліс є потужним ансамблевим методом, що поєднує в собі велику кількість рішень дерев. Завдяки цьому він добре справляється зі складними нелінійними залежностями, демонструє високу точність і стійкість до перенавчання. Його головним недоліком є менша прозорість порівняно з логістичною регресією, хоча сучасні методи візуалізації важливості ознак частково компенсують цю проблему.

З позиції використання в задачах медичного прогнозування, всі три моделі мають свої обґрунтовані сфери застосування. Логістична регресія підходить для швидкої оцінки ризику та побудови базових прототипів. Метод KNN корисний у випадках, коли є добре збалансовані й стандартизовані набори даних, однак його обчислювальна вартість ускладнює масштабування. Випадковий ліс забезпечує кращу загальну продуктивність при роботі з високовимірними або нерівномірно розподіленими даними.

Таким чином, вибір моделі для передбачення серцевого нападу залежить від доступних ресурсів, вимог до інтерпретованості, а також конкретних характеристик даних. У рамках цієї роботи буде проведено емпіричне порівняння згаданих методів на основі відкритого набору даних, що містить медичні показники пацієнтів, із метою визначення найефективнішого підходу до виявлення ризику серцевого нападу на ранніх стадіях.

3 ТЕОРЕТИЧНЕ ДОСЛІДЖЕННЯ

3.1 Архітектура та проектування ПЗ

Розробка програмного забезпечення для прогнозування ризику серцевого нападу з використанням методів машинного навчання потребує створення модульної, масштабованої та ефективної архітектури. Така система повинна не лише забезпечити високу точність прогнозу, а й бути адаптованою до медичних умов, зокрема до обмежень за обсягом та якістю даних, а також вимог щодо прозорості рішень.

Архітектура ПЗ умовно поділяється на п’ять основних модулів: модуль збору та підготовки даних, модуль побудови та навчання моделей, модуль прогнозування, модуль візуалізації результатів та модуль зберігання даних.

Модуль збору та підготовки даних відповідає за обробку медичних даних пацієнтів, таких як артеріальний тиск, рівень холестерину, електрокардіограма (ЕКГ), наявність стенокардії, вік, стать, частота серцевих скорочень та інші клінічні ознаки. Дані очищуються від пропусків і аномалій, масштабуються, кодуються та розділяються на навчальну і тестову вибірки. У реалізації на Python застосовуються бібліотеки pandas, scikit-learn, numpy та seaborn для аналізу і підготовки даних.

Модуль побудови моделей реалізує машинне навчання для класифікації пацієнтів за ризиком серцевого нападу. У рамках дослідження реалізовано та порівняно кілька моделей: логістичну регресію, метод k-найближчих сусідів (KNN) та випадковий ліс (Random Forest). Для реалізації моделей використовуються бібліотеки scikit-learn, xgboost та TensorFlow/Keras. Застосовується перехресна валідація (cross_val_score) для об’єктивної оцінки якості моделей та пошуку оптимальних гіперпараметрів (GridSearchCV, RandomizedSearchCV).

Модуль прогнозування відповідає за застосування навчених моделей до нових (раніше невідомих) даних. Вхідними є медичні показники нового пацієнта, а вихідним – ймовірність серцевого нападу або відповідний клас ризику. У

системі реалізовані механізми перевірки на коректність введення та інтерпретація результатів.

Модуль візуалізації результатів дозволяє лікарю або медичному фахівцю переглядати результати прогнозу в зручній формі. Для цього створено веб-інтерфейс за допомогою Streamlit, що дозволяє завантажити дані пацієнта, переглянути прогнози, графіки розподілу ризику та інші статистичні метрики. Графіки точності, ROC-криві, матриці неточностей генеруються засобами matplotlib та seaborn.

Модуль зберігання даних призначений для ведення обліку результатів класифікації та історії пацієнтів. У повноцінному рішенні можливе підключення до бази даних. У прототипі реалізовано збереження результатів у .csv або .json форматах для подальшого аналізу. Така архітектура дозволяє зберігати та порівнювати історичні результати, проводити аудит прийнятих моделей та виявляти патерни ризику.

Розроблене ПЗ дозволяє лікарям ефективно і швидко оцінювати ризики на основі як класичних моделей, так і сучасних глибоких методів. Завдяки модульній структурі система легко адаптується до розширення – наприклад, для включення нових ознак або моделей, інтеграції з медичними інформаційними системами, або переходу на режим реального часу.

3.2 Обмеження у процесі дослідження

У процесі розробки програмної системи для передбачення серцевого нападу з використанням методів машинного навчання виникає низка обмежень, які слід враховувати при побудові, тренуванні та оцінюванні моделей.

Обмеження якості та доступності медичних даних. Для ефективного навчання моделей машинного навчання необхідні великі, якісні та збалансовані набори медичних даних. Проте реальні медичні дані часто містять пропуски, шум, дублікати або нерепрезентативні приклади. Крім того, доступ до клінічних записів обмежується етичними, юридичними та конфіденційними аспектами. Ці

чинники можуть впливати на якість навчання моделей та їх здатність до узагальнення.

Нерівномірність класів. Проблема дисбалансу між кількістю пацієнтів, у яких був зафіксований серцевий напад, і тих, у кого напад не відбувся, є типовою для медичних датасетів. Така ситуація ускладнює навчання моделей, які можуть зміщуватися в бік переважного класу, і потребує використання спеціальних методів, таких як збалансовані метрики (AUC-ROC, F1-score), oversampling або undersampling, або ж модифіковані функції втрат.

Гетерогенність медичних ознак. Ознаки, що використовуються у прогнозуванні (вік, тиск, рівень холестерину, наявність болю, ЕКГ тощо), можуть суттєво варіюватися залежно від пацієнтів, медичних центрів і навіть способу збору даних. Такі відмінності можуть викликати складності при генералізації моделей на нові набори даних.

Відсутність пояснюваності моделей. Багато потужних моделей (наприклад, ансамблеві або нейронні мережі) функціонують як "чорні ящики", що ускладнює інтерпретацію результатів. У медичному контексті критично важливо розуміти причини класифікації, тому варто впроваджувати додаткові засоби пояснення рішень, такі як SHAP або LIME.

Труднощі з адаптацією до нових даних. Дані, зібрані в одній медичній установі, можуть суттєво відрізнятися від даних з іншого джерела, що створює проблему переносу (transferability) моделі. Для вирішення цієї проблеми може знадобитися донавчання моделей або застосування методів доменної адаптації.

Обмеження обчислювальних ресурсів. Деякі моделі (наприклад, ансамблеві або глибокі нейронні мережі) можуть вимагати значних обчислювальних ресурсів, особливо на етапі навчання. Це може бути критично при обмеженому апаратному забезпеченні або при необхідності обробки в реальному часі.

Таким чином, успішна реалізація системи передбачення серцевого нападу вимагає врахування зазначених обмежень та застосування відповідних методів їх компенсації. В подальшому перспективним напрямком є використання комбінованих моделей, які поєднують точність, інтерпретованість і адаптивність.

3.3 Необхідні ресурси для виконання дослідження

Для ефективного проведення дослідження методів машинного навчання, спрямованих на передбачення серцевого нападу, необхідно забезпечити відповідні технічні, інформаційні та програмні ресурси. Всі ресурси умовно можна поділити на три категорії: технічні засоби, набори даних і ресурси для навчання та тестування моделей.

Технічні ресурси:

- персональний комп’ютер із сучасним багатоядерним процесором (CPU) та бажано з графічним процесором (GPU), який суттєво пришвидшує тренування моделей, особливо при використанні ансамблевих або нейронних підходів;
- середовище програмування Python та інтегроване середовище розробки, для зручності розробки та тестування моделей;
- бібліотеки машинного навчання, зокрема scikit-learn, XGBoost, LightGBM, а також TensorFlow або Keras у разі використання глибокого навчання;
- системи для обробки та візуалізації даних, як-от pandas, NumPy, Matplotlib, Seaborn.

Інформаційні ресурси (набори даних):

- наявність якісного датасету медичних показників пацієнтів, наприклад, набору Heart Disease UCI, Cleveland Heart Disease dataset або Framingham Heart Study, які містять такі ознаки, як артеріальний тиск, рівень холестерину, частота серцевих скорочень, вік, стать, історія хвороб тощо;
- можливість попередньої обробки даних, включаючи заповнення пропусків, масштабування ознак, перетворення категоріальних змінних та побудову нових ознак (feature engineering).

Ресурси для тренування, оцінювання і тестування моделей:

- інструменти для розподілу вибірки на тренувальну, валідаційну та тестову частини, що реалізується за допомогою функцій train_test_split з scikit-learn;
- метрики якості моделей: accuracy, precision, recall, F1-score, ROC AUC, що дозволяють об’єктивно оцінити ефективність прогнозування серцевих нападів;
- можливість гіперпараметричної оптимізації моделей з використанням GridSearchCV або RandomizedSearchCV для підвищення точності;
- візуалізація результатів – побудова кривих ROC, confusion matrix, діаграм важливості ознак тощо.

Забезпечення перелічених ресурсів дозволить побудувати ефективну, надійну та інтерпретовану модель машинного навчання, яка може слугувати основою для створення прикладної системи раннього попередження про ризик серцевого нападу.

3.4 Проектування системи

Розробка системи для передбачення серцевого нападу базується на модульному підході, який забезпечує чіткий розподіл функцій між компонентами, гнучкість у налаштуванні системи та можливість масштабування. Такий підхід дозволяє незалежно оновлювати та оптимізувати окремі модулі без впливу на інші частини системи. Система повинна бути розроблена таким чином, щоб забезпечити високу точність прогнозування, інтерпретованість результатів і можливість інтеграції з існуючими медичними інформаційними системами.

Основні компоненти системи включають:

- модуль збору та підготовки даних – відповідає за завантаження, очищення, нормалізацію медичних показників та формування вибірки для навчання;
- модуль машинного навчання – реалізує побудову, навчання та збереження моделей машинного навчання. Залежно від обраного

підходу, це можуть бути логістична регресія, Random Forest, XGBoost або нейронна мережа;

– модуль оцінювання результатів – забезпечує обчислення метрик точності, візуалізацію кривих ROC, confusion matrix та інтерпретацію результатів на основі важливості ознак;
– інтерфейс користувача – призначений для взаємодії з медичним працівником, відображення прогнозів та основних характеристик пацієнта.

Для наочного представлення структури системи та взаємодії між її елементами наведено use-case діаграму. Серед основних акторів:

– пацієнт – особа, дані якої обробляються системою для прогнозу ризику серцевого нападу;
– медичний працівник – взаємодіє з системою, переглядає результати прогнозу та приймає клінічні рішення;
– система прогнозування – аналізує медичні показники та здійснює прогнозування;
– система моніторингу даних – зовнішній компонент, який надає або приймає дані.

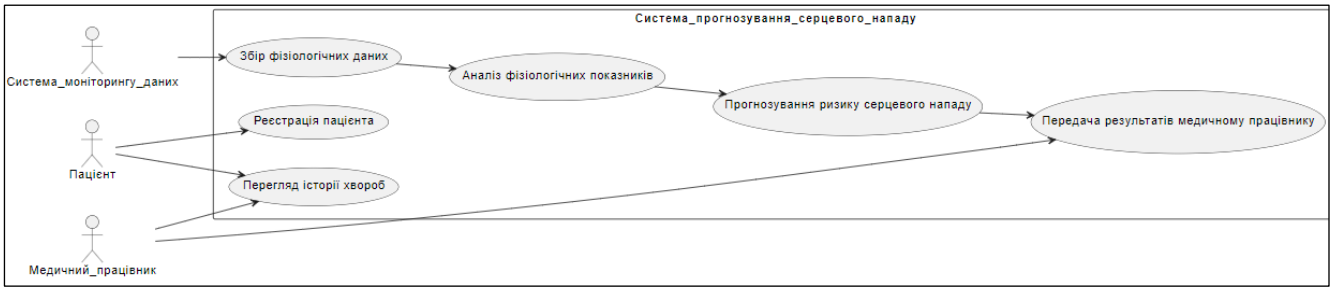

Рисунок 3.4.1 – Use-case діаграма системи (рисунок виконано самостійно)

Ключовою вимогою до проектування є можливість інтеграції з електронними медичними записами, лабораторними базами даних та іншими системами охорони здоров’я. Це дозволяє отримувати актуальну інформацію про стан пацієнта, автоматично оновлювати модель новими даними та формувати персоналізовані прогнози ризику.

Таким чином, проектована система має бути не лише інструментом для аналізу даних, а й ефективним засобом підтримки клінічних рішень у медичній практиці.

3.5 UI/UX Дизайн Системи

Дизайн користувацького інтерфейсу (UI) системи для прогнозування серцевого нападу має бути максимально зрозумілим, функціональним та зосередженим на підтримці прийняття клінічних рішень [10]. Основними користувачами системи є лікарі-кардіологи, медичний персонал та аналітики. Основна мета UI-дизайну – надати швидкий доступ до ключової інформації про стан пацієнта, результати моделювання та історію медичних показників.

Система передбачає один основний тип інтерфейсу – інтерфейс для медичного працівника, який реалізується у вигляді веб-додатку. До основних його елементів належать:

- панель управління з узагальненими показниками ризику серцевого нападу для кожного пацієнта;
- інтерактивні графіки для візуалізації динаміки змін медичних параметрів (наприклад, рівень холестерину, артеріальний тиск, пульс тощо);
- вивід результатів прогнозування у вигляді ймовірності настання серцевого нападу у найближчий період;
- форма введення нових даних;
- модуль порівняння моделей, що дозволяє лікарю обрати кращу модель за метриками точності, recall або precision.

Інтерфейс орієнтований на мінімалістичний стиль: чітка типографіка, акценти на важливих медичних індикаторах. Це дозволяє знизити когнітивне навантаження на користувача та полегшити візуальне сприйняття.

UX-дизайн системи враховує такі аспекти:

- плавна навігація між пацієнтами, результатами моделей та історіями спостережень;

- можливість швидкого імпорту або підключення до електронної медичної картки пацієнта;
- адаптивність інтерфейсу для роботи як на настільних комп’ютерах, так і на планшетах та мобільних пристроях.

Таким чином, UI/UX-дизайн системи має забезпечити лікарю простий, інтуїтивний інструмент для прийняття рішень, що базуються на результатах машинного навчання, з акцентом на швидкість взаємодії, інтерпретованість та точність.

3.6 Приклади алгоритмів та методів

У процесі розробки системи для прогнозування серцевого нападу важливим етапом є вибір ефективних алгоритмів машинного навчання, а також методів попередньої обробки та аналізу медичних даних. Нижче наведено приклади алгоритмів та їх застосування в контексті дослідження.

До початку навчання моделей необхідно здійснити очистку й трансформацію вхідних медичних даних:

- видалення або імпутація пропущених значень, усунення викидів;
- застосування MinMaxScaler або StandardScaler для приведення даних до одного масштабу, що особливо важливо для моделей, чутливих до масштабу (наприклад, KNN);
- застосування One-Hot Encoding або Label Encoding для перетворення категорійних змінних у числові вектори;
- балансування класів: використання методів, таких як SMOTE (Synthetic Minority Over-sampling Technique) для збалансування кількості прикладів у кожному класі (особливо якщо дані є несбалансованими).

Для реалізації основної задачі прогнозування серцевого нападу можуть використовуватися такі алгоритми:

- логістична регресія – проста інтерпретована модель, яка оцінює ймовірність настання події (серцевого нападу) на основі медичних параметрів;

- random forest – ансамблевий метод, що використовує багато дерев рішень для підвищення точності та стійкості до шуму;
- k-nearest neighbors (KNN) – класифікатор, що приймає рішення на основі схожості нових даних з уже відомими прикладами.

Оптимізація та налаштування моделей:

- гіперпараметрична оптимізація: використання Grid Search або Randomized Search для підбору найкращих параметрів моделей (наприклад, кількість дерев у Random Forest, коефіцієнт регуляризації в логістичній регресії тощо);
- застосування L1/L2-регуляризації для запобігання перенавчанню (overfitting), особливо для логістичної регресії;
- перехресна перевірка (Cross-validation): оцінювання якості моделі на декількох підмножинах даних, що дозволяє уникнути переоцінки точності.

Метрики якості моделі

Оскільки мова йде про медичну діагностику, особливу увагу приділено не тільки точності, але й:

- Recall (повнота) – важливо виявити всі справжні випадки серцевого нападу;
- Precision (точність) – мінімізувати кількість хибнопозитивних результатів;
- F1-score – баланс між recall і precision;
- ROC AUC – загальна міра здатності моделі розділяти класи.

3.8 Реалізація алгоритму визначення хвороби

У рамках розробки системи для передбачення серцевого нападу ключовим етапом є побудова, навчання та оцінка моделі машинного навчання, здатної аналізувати медичні дані пацієнта та робити прогноз ризику виникнення серцевого нападу. Така система повинна забезпечувати високий рівень точності, інтерпретованості та адаптивності до різних вхідних даних.

Підготовка та обробка даних. Початковим етапом є збирання відповідного набору даних. Для цього використовується медичний датасет, який містить записи пацієнтів із відповідними ознаками (наприклад, вік, стать, артеріальний тиск, рівень холестерину, серцевий ритм тощо) та мітками про наявність/відсутність серцевого нападу.

Після завантаження даних виконується їхня очистка, зокрема:

- заповнення пропущених значень або їх видалення;
- нормалізація числових ознак (масштабування у діапазон 0–1 або Z-нормалізація);
- кодування категоріальних змінних (наприклад, OneHot або Label Encoding).

Розробка моделі.Для вирішення задачі бінарної класифікації застосовуються моделі машинного навчання, такі як логістична регресія, метод k-ближчих сусідів або випадковий ліс. Кожна модель має свої переваги, і їх ефективність порівнюється під час навчання.

Також виконується пошук найкращих гіперпараметрів за допомогою GridSearchCV або RandomizedSearchCV. Для підвищення достовірності моделі застосовується перехресна перевірка (k-fold cross-validation).

Фрагмент коду навчання моделі:

```
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
import pandas as pd
import joblib

# Завантаження та попередня обробка даних
df = pd.read_csv("heart_disease_data.csv")
X = df.drop("target", axis=1)
y = df["target"]

# Розділення на тренувальні та тестові дані
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Побудова моделі та гіперпараметричний пошук
param_grid = {
    "n_estimators": [50, 100, 150],
    "max_depth": [3, 5, 7],
```

```
    "min_samples_split": [2, 5],
}

model = GridSearchCV(RandomForestClassifier(random_state=42),
param_grid, cv=5)
model.fit(X_train, y_train)

# Оцінка моделі
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

Оцінка результатів. Якість моделі визначається за такими метриками: точність (accuracy), повнота (recall), F1-міра та площа під ROC-кривою. Особлива увага приділяється мінімізації кількості хибнонегативних результатів, оскільки у випадку серцевого нападу помилкове "здоровий" може мати фатальні наслідки.

Інтерактивна реалізація. Для практичного використання моделі створюється Streamlit-застосунок, який дозволяє лікарю або користувачу вводити параметри пацієнта (вік, тиск тощо) та отримувати прогноз ризику серцевого нападу в реальному часі:

```
import streamlit as st
import joblib
import numpy as np

st.title("Прогноз ризику серцевого нападу")

# Завантаження моделі
model = joblib.load("best_model.pkl")

# Інтерфейс введення даних
age = st.slider("Вік", 20, 90, 50)
chol = st.slider("Холестерин", 100, 400, 200)
trestbps = st.slider("Тиск", 90, 180, 120)
thalach = st.slider("Макс. серцевий ритм", 60, 200, 150)
oldpeak = st.slider("Depression", 0.0, 6.0, 1.0)

# Масив ознак (спрощено)
input_data = np.array([[age, chol, trestbps, thalach, oldpeak]])

# Прогноз
if st.button("Передбачити"):
    prediction = model.predict(input_data)[0]
    prob = model.predict_proba(input_data)[0][prediction]

    if prediction == 1:
        st.error(f"Високий ризик серцевого нападу ({prob:.2%})")
    else:
        st.success(f"Низький ризик серцевого нападу ({prob:.2%})")
```

Ця реалізація охоплює повний цикл створення інтелектуальної системи для передбачення серцевого нападу: від збору та обробки даних, до побудови моделі, її оцінки.

4 ПРАКТИЧНЕ ДОСЛІДЖЕННЯ

Для дослідження можливостей методів машинного навчання щодо передбачення серцевого нападу було реалізовано програмний застосунок, що дозволяє виконувати попередню обробку даних, тренування моделей, гіперпараметричний пошук та оцінку результатів класифікації. У якості вхідних даних використано датасет Heart Disease UCI Dataset, що містить інформацію про фізіологічні показники пацієнтів.

Початковий інтерфейс застосунку реалізовано за допомогою бібліотеки Streamlit, з можливістю візуалізації результатів (див. рис. 4.1).

Рисунок 4.1 – Інтерфейс застосунку (рисунок виконано самостійно)

В першу чергу необхідно провести аналіз наших даних для визначення того чи підходить датасет для тренування моделей. Почнемо з перевірки розподілу значень нашої цільової змінної (target), яка відповідає за те чи хвора людина, для визначення збалансованості обраного датасету (див .рис. 4.2).

Оскільки кількість позитивних та негативних зразків близька до рівної, наш цільовий стовпець можна вважати збалансованим. Незбалансований цільовий стовпець, де деякі класи мають значно більше зразків, може бути складнішим для моделювання, ніж збалансований набір. В ідеалі, всі цільові класи повинні мати однакову кількість зразків.

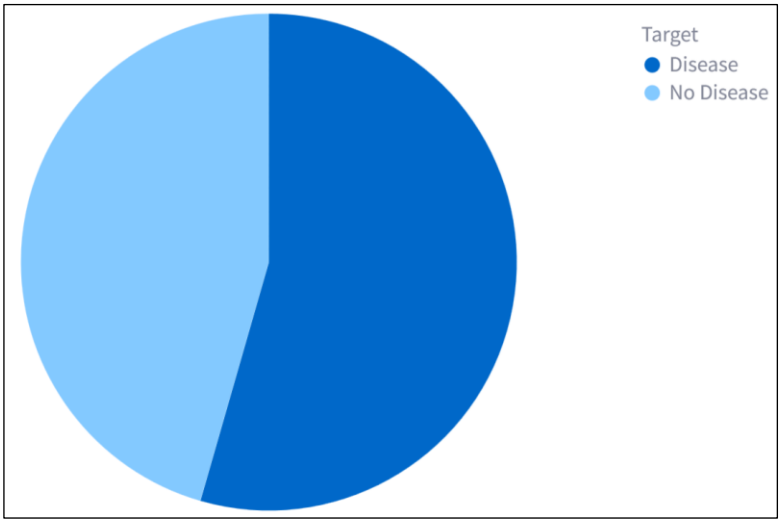

Рисунок 4.2 – Розподіл значень нашої цільової змінної (рисунок виконано самостійно)

Далі проаналізуємо частоту хвороб серця відповідно до статі (див. рис 4.3)

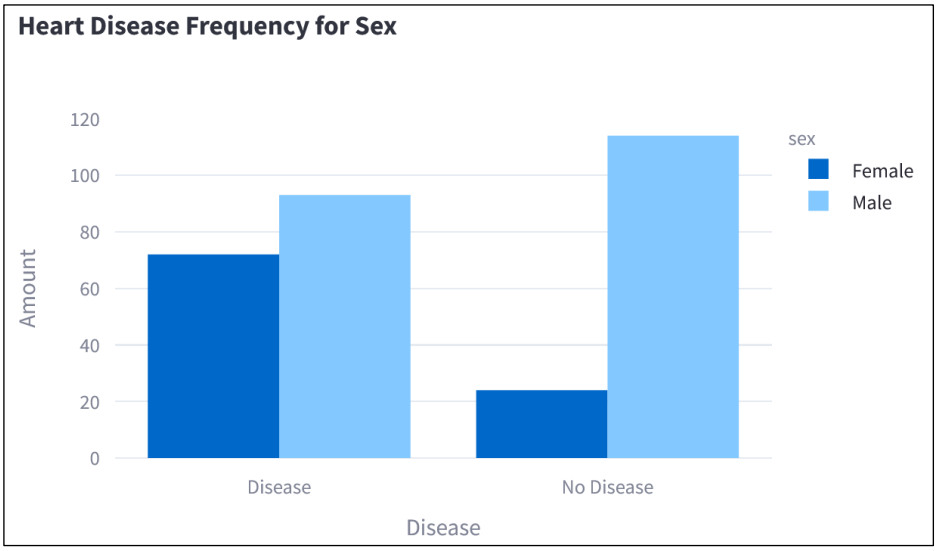

Рисунок 4.3 – Частота хвороб серця за статтю (рисунок виконано самостійно)

Згідно з даним графіком, для жінок з приблизно 100 осіб близько 72 мають позитивне значення, що вказує на наявність серцевих захворювань. Таким чином, виходячи виключно з статі, існує приблизно 75% ймовірність серцевих захворювань.

Для чоловіків з приблизно 200 осіб приблизно половина вказує на наявність серцевих захворювань. Таким чином, виходячи виключно з статі, існує приблизно 50% ймовірність серцевих захворювань.

Усереднюючи ці два значення, ми можемо встановити просту евристику: якщо ми більше нічого не знаємо про людину, існує 62,5% ймовірність того, що у неї є серцеві захворювання.

Ця базова евристика може служити нашою початковою базовою лінією, яку ми прагнемо перевершити за допомогою методів машинного навчання.

Тепер щоб дослідити зв'язок між віком і максимальною частотою серцевих скорочень (thalach) щодо серцевих захворювань, ми створимо діаграму розсіювання. Ця діаграма дозволить нам візуалізувати розподіл точок даних і виявити потенційні закономірності або тенденції.

Наступним кроком, щоб дослідити зв'язок між віком і максимальною частотою серцевих скорочень щодо серцевих захворювань, ми створимо діаграму розсіювання (див. рис 4.4). Ця діаграма дозволить нам візуалізувати розподіл точок даних і виявити потенційні закономірності або тенденції.

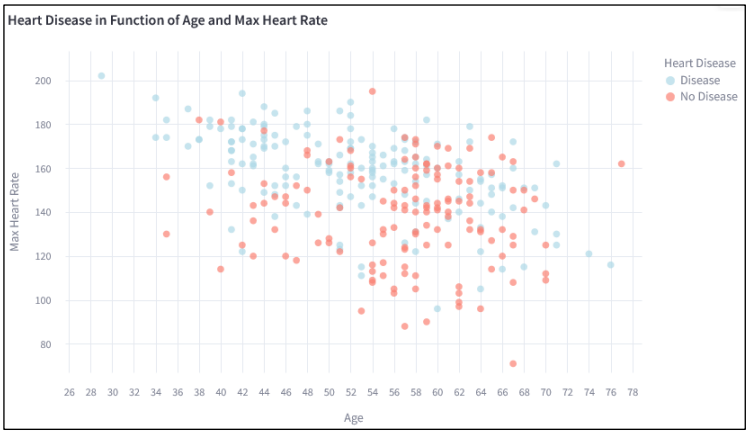

Рисунок 4.4 – Діаграма розсіювання (рисунок виконано самостійно)

Спостереження показують, що молодші люди, як правило, мають вищу максимальну частоту серцевих скорочень, про що свідчить розташування точок вище в лівій частині графіка. І навпаки, серед старших учасників спостерігається переважання синіх точок (що свідчать про наявність серцевих захворювань), що потенційно може бути пов'язано з більшою концентрацією точок у правій частині графіка (що представляють старших людей).

Ці спостереження дають початкове уявлення про дані, сприяючи нашому розумінню їх характеристик.

Гістограми є корисним методом оцінки розподілу змінної, створимо одну для візуалізації розподілу віку (див. рис. 4.5).

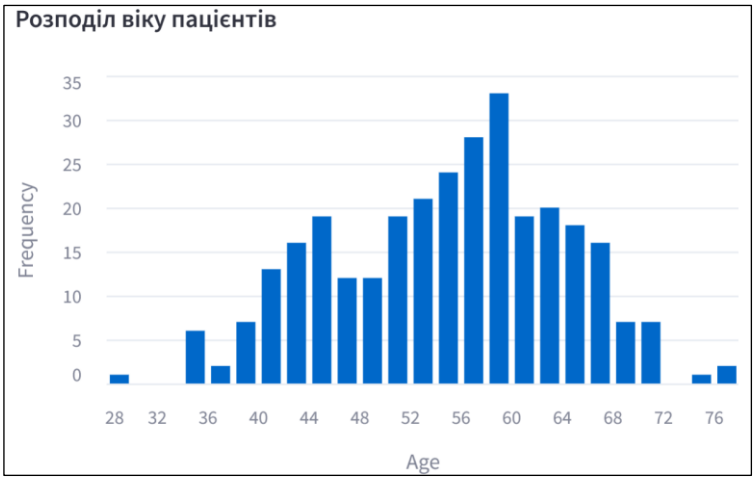

Рисунок 4.5 – Розподіл значень віку (рисунок виконано самостійно)

Розподіл за віком, схоже, відповідає дещо праворуч зміщеному нормальному розподілу, що підтверджується гістограмою. Ця характеристика відображається на діаграмі розсіювання, сприяючи спостережуваним закономірностям.

Ми продовжимо дослідженням іншої незалежної змінної, а саме біль у грудях (ср), використовуючи подібний підхід до нашого аналізу зі статтю (див. рис 4.6).

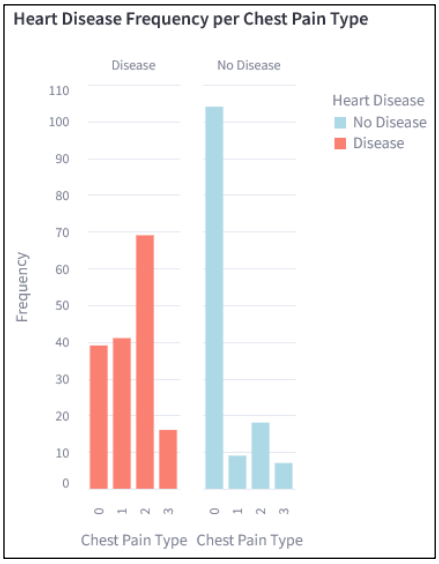

Рисунок 4.6 – Частота хвороби в залежності від типу болю (рисунок виконано самостійно)

Щоб отримати уявлення про взаємозв'язки між усіма незалежними змінними та нашою цільовою змінною, ми порівняємо їх за допомогою кореляційної матриці (див. рис. 4.7).

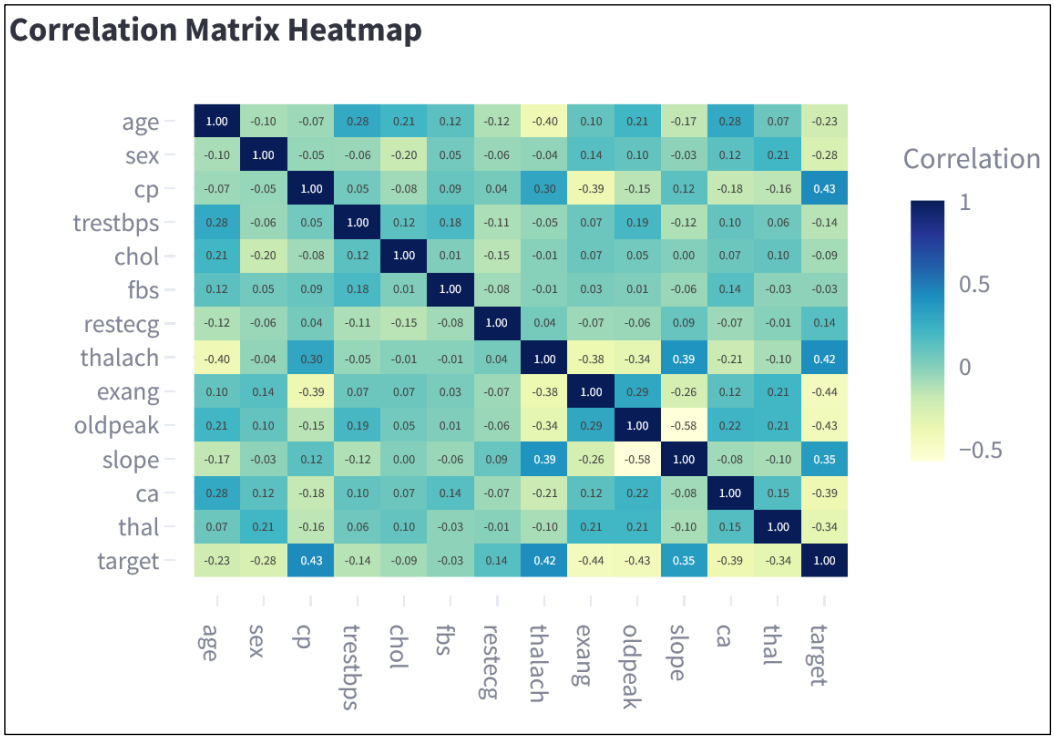

Рисунок 4.7 – Кореляційна матриця (рисунок виконано самостійно)

Кореляційна матриця надає таблицю числових значень, що вказують на ступінь зв'язку між кожною парою змінних. Цей аналіз може допомогти визначити, які незалежні змінні потенційно можуть впливати на нашу цільову змінну.

Провівши розвідувальний аналіз, ми почали розуміти наш набір даних. Окрім нашої базової оцінки з використанням статі, решта даних, здається, розподілені справедливо. Далі ми перейдемо до модельно-орієнтованого розвідувального аналізу, використовуючи моделі машинного навчання для керівництва нашими подальшими дослідженнями.

Давайте приступимо до створення моделей. Дослідивши дані, ми тепер використаємо машинне навчання для прогнозування нашої цільової змінної на основі 13 незалежних змінних. Перш ніж створювати модель, нам потрібно підготувати наш набір даних.

Ми прагнемо передбачити нашу цільову змінну (target), використовуючи всі інші змінні. Для досягнення цього нам потрібно відокремити цільову змінну (target) від решти набору даних.

```
# Everything except target variable
X = df.drop("target", axis=1)

# Target variable
y = df.target.values
```

Однією з фундаментальних концепцій машинного навчання є розділення на навчальний та тестовий набір. Навчальний набір використовується для навчання вашої моделі, а тестовий – для оцінки її продуктивності.

Тестовий набір максимально точно імітує розгортання вашої моделі в реальному сценарії. Вкрай важливо ніколи не дозволяти вашій моделі навчатися на тестовому наборі; її слід оцінювати лише на ньому.

Щоб розділити наші дані на навчальний та тестовий набори, ми можемо використовувати функцію `train_test_split()` Scikit-Learn, передаючи їй наші незалежні та залежні змінні (X та y).

Параметр `test_size` у функції `train_test_split()` визначає частку наших даних, виділених для тестового набору. Поширеною практикою є виділення 80% даних для навчання, а решта 20% для тестування.

Після підготовки даних ми готові до апроксимації моделей. Ми використовуватимемо такі алгоритми та порівнюватимемо їхні результати:

- логістична регресія;
- k-найближчих сусідів;
- випадковий ліс.

Бібліотека Scikit-Learn пропонує узгоджений інтерфейс для всіх алгоритмів. Як навчання моделі (`model.fit(X_train, y_train)`), так і її оцінювання (`model.score(X_test, y_test)`) дотримуються тих самих правил. Метод `score()` повертає співвідношення правильних прогнозів (1.0 вказує на 100% точність).

Після тренування і оцінки наших моделей, ми можемо візуалізувати і порівняти їх точність (див. рис 4.8).

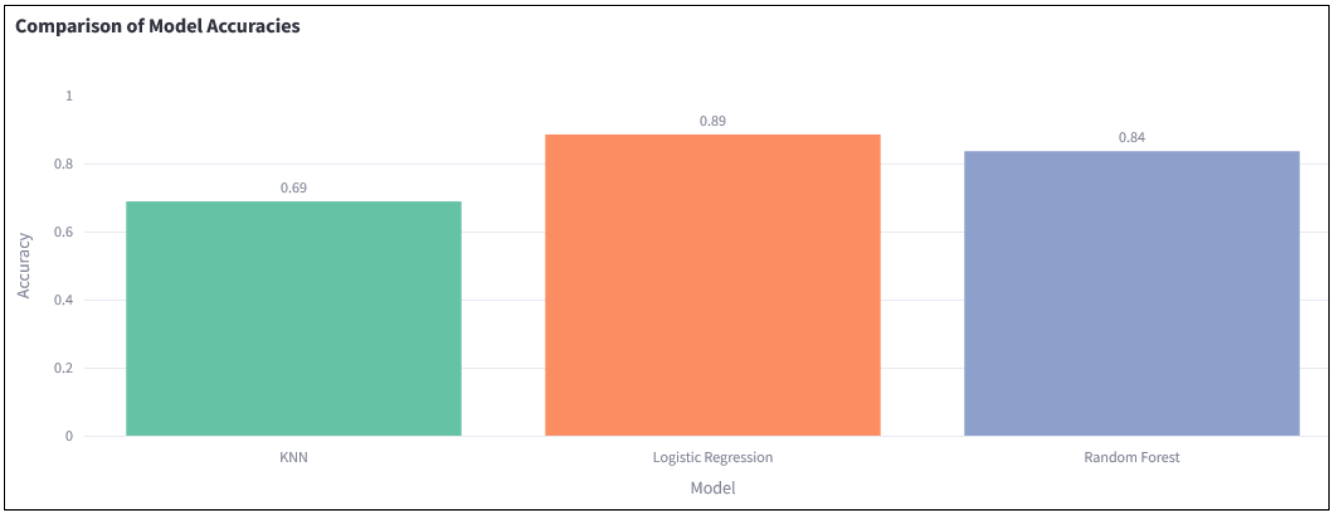

Рисунок 4.8 – Порівняння моделей (рисунок виконано самостійно)

Ми можемо створити звіт за допомогою функції `classification_report()`, яка приймає справжні та передбачені дані у якості вхідних. Цей звіт надає інформацію про точність, повноту та F1-оцінку нашої моделі для кожного класу. Створимо звіт для кожної з моделей (див. рис. 4.9, 4.10, 4.11).

	precision	recall	f1-score	support
0	0.6923	0.6207	0.6545	29
1	0.6857	0.75	0.7164	32
accuracy	0.6885	0.6885	0.6885	0.6885
macro avg	0.689	0.6853	0.6855	61
weighted avg	0.6888	0.6885	0.687	61

Рисунок 4.9 – Звіт до KNN моделі (рисунок виконано самостійно)

	precision	recall	f1-score	support
0	0.8929	0.8621	0.8772	29
1	0.8788	0.9063	0.8923	32
accuracy	0.8852	0.8852	0.8852	0.8852
macro avg	0.8858	0.8842	0.8848	61
weighted avg	0.8855	0.8852	0.8851	61

Рисунок 4.10 – Звіт до моделі логістичної регресії (рисунок виконано самостійно)

	precision	recall	f1-score	support
0	0.8276	0.8276	0.8276	29
1	0.8438	0.8438	0.8438	32
accuracy	0.8361	0.8361	0.8361	0.8361
macro avg	0.8357	0.8357	0.8357	61
weighted avg	0.8361	0.8361	0.8361	61

Рисунок 4.11 – Звіт до моделі випадкового лісу (рисунок виконано самостійно)

В даних звітах метрики означають наступне:

– precision: вказує частку позитивних ідентифікацій (модель передбачила клас 1), які були фактично правильними. Модель без хибнопозитивних результатів має точність 1.0;
– recall: вказує частку фактичних позитивних результатів, які були правильно класифіковані. Модель без хибнонегативних результатів має повність 1.0;
– f1-score: поєднання точності та повності. Ідеальна модель досягає оцінки F1 1.0;
– support: кількість зразків, на яких розраховувався кожен показник;
– accuracy: точність моделі в десятковій формі. Ідеальна точність дорівнює 1.0;
– macro avg: скорочення від макросереднє, це середнє значення точності, повності та оцінки F1 між класами. Воно не враховує дисбаланс класів, тому важливо звертати увагу на цей показник, якщо є дисбаланси класів;
– weighteg avg: скорочення від середньозважене значення, це середньозважене значення точності, повноти та F1-балу між класами. Кожен показник розраховується на основі кількості зразків у кожному класі. Цей показник надає перевагу класу-мажору та дає високе значення, коли один клас перевершує інший завдяки більшій кількості зразків.

Тепер перейдемо до оптимізації гіперпараметрів і перехресної валідації. З огляду на обмежені дані, ми використовуватимемо перехресну валідацію, а не окремий набір для перевірки, для експериментів з гіперпараметрами.

Найпоширенішим підходом є k-кратна перехресна перевірка, де дані розділяються на k підмножин для тестування. Наприклад, з 5 множинами (k = 5) кожна множина представляє сегмент даних, що використовуються для тестування в ротації.

Почнемо з налаштування гіперпараметрів алгоритму K-найближчих сусідів (KNN). Для KNN основним гіперпараметром, який ми можемо налаштувати, є кількість сусідів. Оцінимо ефективність моделі при кількості сусідів від 1 до 20 (див. рис. 4.12).

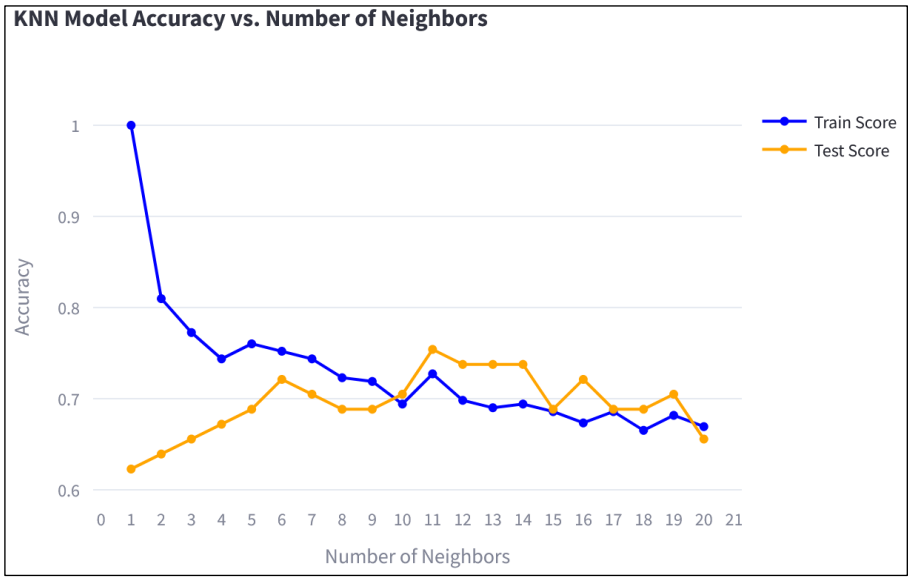

Рисунок 4.12 – Точність моделі KNN від кількості сусідів (рисунок виконано самостійно)

Дивлячись на графік, видно, що 11 сусідів забезпечують найкращу продуктивність моделі. Виведемо звіт для оцінки налаштованої моделі (див. рис. 4.13).

	precision	recall	f1-score	support
0	0.7692	0.6897	0.7273	29
1	0.7429	0.8125	0.7761	32
accuracy	0.7541	0.7541	0.7541	0.7541
macro avg	0.756	0.7511	0.7517	61
weighted avg	0.7554	0.7541	0.7529	61

Рисунок 4.13 – Звіт до KNN моделі після налаштування (рисунок виконано самостійно)

В результаті налаштування гіперпараметрів нам вдалося покращити точність моделі на 7%. Однак, навіть з урахуванням цього, продуктивність моделі K-найближчих сусідів (KNN) не наблизилася до продуктивності логістичної регресії або випадкового лісу. Враховуючи це, ми відкинемо KNN і зосередимося на двох інших моделях.

Замість ручного налаштування логістичної регресії та випадкового лісу, давайте використаємо RandomizedSearchCV. RandomizedSearchCV автоматично пробує кілька різних комбінацій гіперпараметрів, оцінює їх і вибирає найкращу.

Ми надамо йому різні гіперпараметри з попередньо створеного словника та встановимо кількість ітерацій на 20. Це означає, що RandomizedSearchCV дослідить 20 різних комбінацій гіперпараметрів зі словника та вибере найкращі з них.

Тепер, коли ми налаштували логістичну регресію за допомогою RandomizedSearchCV, давайте зробимо те саме для випадкового лісу, після чого переглянемо звіти налаштованих моделей (див. рис. 4.14, 4.15)

	precision	recall	f1-score	support
0	0.8889	0.8276	0.8571	29
1	0.8529	0.9063	0.8788	32
accuracy	0.8689	0.8689	0.8689	0.8689
macro avg	0.8709	0.8669	0.868	61
weighted avg	0.87	0.8689	0.8685	61

Рисунок 4.14 – Звіт до моделі випадкового лісу після налаштування (рисунок виконано самостійно)

	precision	recall	f1-score	support
0	0.8929	0.8621	0.8772	29
1	0.8788	0.9063	0.8923	32
accuracy	0.8852	0.8852	0.8852	0.8852
macro avg	0.8858	0.8842	0.8848	61
weighted avg	0.8855	0.8852	0.8851	61

Рисунок 4.15 – Звіт до моделі логістичної регресії після налаштування (рисунок виконано самостійно)

Виходячи з даних звітів, незважаючи на те що нам не вдалося покращити точність моделі логістичної регресії, нам вдалося збільшити точність передбачення випадкового лісу на 3%.

Оскільки логістична регресія показує багатообіцяючі результати, давайте детальніше налаштуємо її за допомогою GridSearchCV. GridSearchCV відрізняється від RandomizedSearchCV тим, що він ретельно тестує кожну можливу комбінацію гіперпараметрів, а не певну кількість ітерацій. Переглянемо отриманий звіт (див. рис. 4.16).

	precision	recall	f1-score	support
0	0.8929	0.8621	0.8772	29
1	0.8788	0.9063	0.8923	32
accuracy	0.8852	0.8852	0.8852	0.8852
macro avg	0.8858	0.8842	0.8848	61
weighted avg	0.8855	0.8852	0.8851	61

Рисунок 4.16 – Звіт до моделі логістичної регресії після подальшого налаштування (рисунок виконано самостійно)

У цьому випадку ми отримуємо ті ж результати, що й раніше, оскільки наша сітка містить максимум 20 різних комбінацій гіперпараметрів. Якщо сітка містить велику кількість комбінацій гіперпараметрів, GridSearchCV може зайняти значний час для оцінки всіх з них. Ось чому часто рекомендується починати з RandomizedSearchCV, щоб дослідити підмножину комбінацій, перш ніж уточнювати пошук за допомогою GridSearchCV.

Далі візуалізуємо ROC-криву та розрахунок показників AUC.

ROC-крива це графічне представлення компромісу між істинно позитивним рівнем (чутливість) та хибнопозитивним рівнем (1 - специфічність) для бінарної моделі класифікації за різних порогових значень. Простіше кажучи, вона допомагає нам зрозуміти, наскільки добре наша модель розрізняє позитивні та негативні класи.

Scikit-Learn надає функцію під назвою `RocCurveDisplay` для створення ROC-кривих та розрахунку метрики площі під кривою (AUC).

Згідно з документацією, ми можемо використовувати метод класу `from_estimator(estimator, X, y)` класу `RocCurveDisplay`, де `estimator` - це підібрана модель машинного навчання, а `X` та `y` - дані, що використовуються для тестування.

У нашому випадку ми використовуватимемо версію GridSearchCV нашого оцінювача логістичної регресії (див. рис. 4.17).

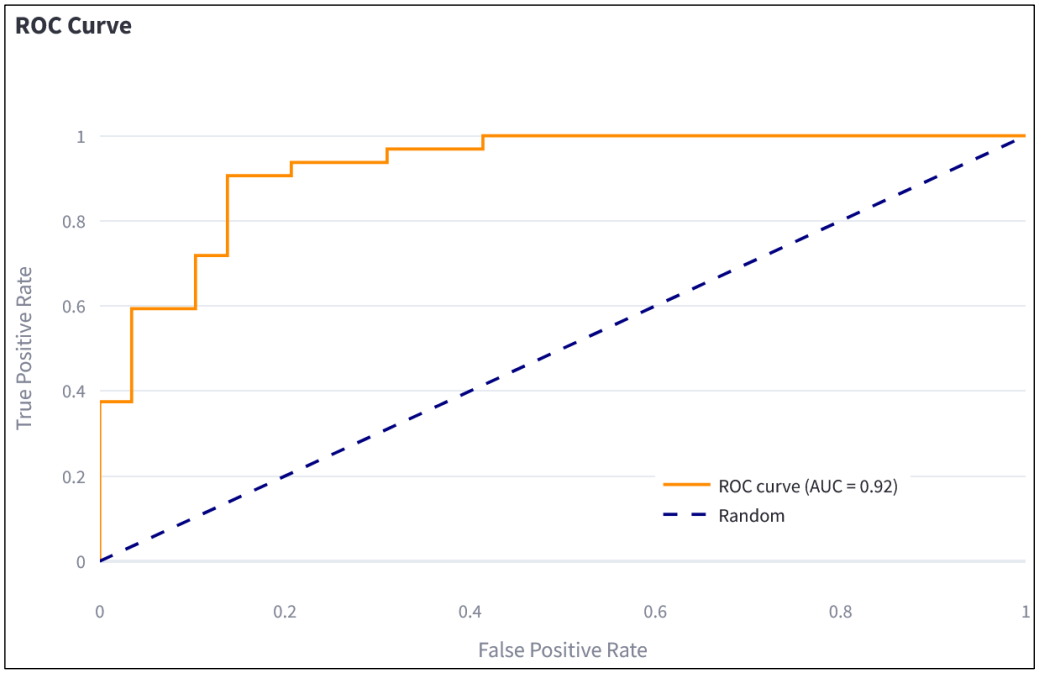

Рисунок 4.17 – ROC крива для логістичної регресії (рисунок виконано самостійно)

Наша модель працює значно краще, ніж випадкове вгадування, яке відповідало б лінії, що йде від нижнього лівого кута до верхнього правого кута (AUC = 0.5). Однак ідеальна модель досягла б показника AUC 1.0, що свідчить про бездоганну продуктивність. Тому все ще є простір для вдосконалення.

Далі розглянемо матрицю плутанини. Матриця плутанини - це візуальне представлення, яке показує, де ваша модель зробила правильні прогнози, а де – неправильні, що дозволяє вам виявити області плутанини. Scikit-Learn надає зручний спосіб створення матриці плутанини за допомогою функції `confusion_matrix()`. Вам просто потрібно передати їй справжні значення та передбачені значення.

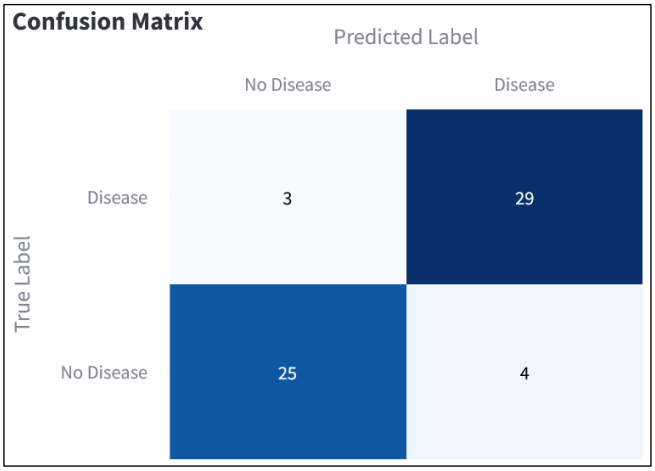

Рисунок 4.18 – Матриця плутанини (рисунок виконано самостійно)

З даної матирці ми можемо спостерігати, що модель допускає однакову кількість помилок (хибнопозитивних та хибнонегативних результатів) в обох класах. Зокрема, є 4 випадки, коли модель передбачала що людина хвора, коли насправді вона здорова (хибнонегативні результати), та 3 випадки, коли модель передбачала здорова замість хворої (хибнопозитивні результати).

Тепер давайте розрахуємо показники зі звітів за допомогою перехресної валідації, щоб зробити їх більш надійними. Ми використовуватимемо найкращу модель разом з найкращими гіперпараметрами та оцінимо їх за допомогою cross_val_score(). Для візуалізації зробимо графік (див. рис. 4.19).

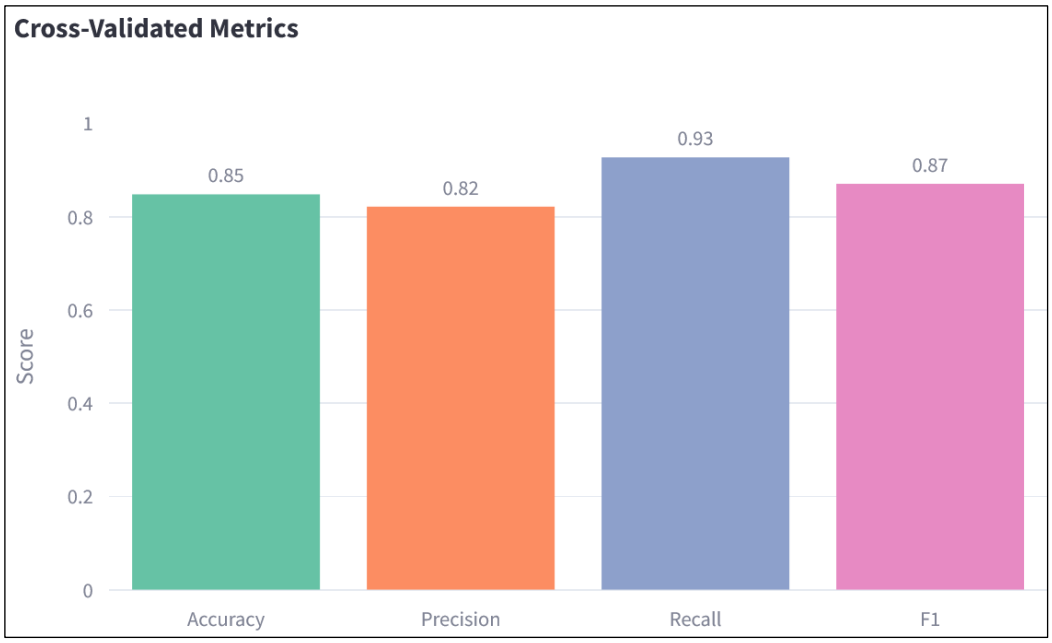

Рисунок 4.19 – Перехресно валідовані метрики (рисунок виконано самостійно)

Остання річ, яку потрібно відмітити у списку наших методів оцінки моделі, це важливість ознак. Важливість ознак - це ще один спосіб дізнатися які ознаки найбільше сприяють результатам моделі. Або для нашої задачі, намагаючись передбачити захворювання серця, використовуючи медичні характеристики пацієнта, які характеристики найбільше впливають на прогнозування.

На відміну від деяких інших функцій, які ми розглядали, оскільки те, як кожна модель знаходить закономірності в даних, дещо відрізняється, те, як модель оцінює важливість цих закономірностей, також відрізняється. Це означає, що для кожної моделі існує дещо інший спосіб визначення того, які ознаки були найважливішими.

Зазвичай ви можете знайти приклад у документації Scikit-Learn. Оскільки ми використовуємо логістичну регресію, ми розглянемо важливості ознак для неї.

Ми можемо створити стовпчасту діаграму, щоб візуалізувати, як кожна ознака впливає на процес прийняття рішень моделі. Вісь x представлятиме назви ознак, а вісь y – коефіцієнти. Це дасть чітке розуміння того, які ознаки мають найбільший вплив на прогнозування серцевих захворювань (див. рис. 4.20).

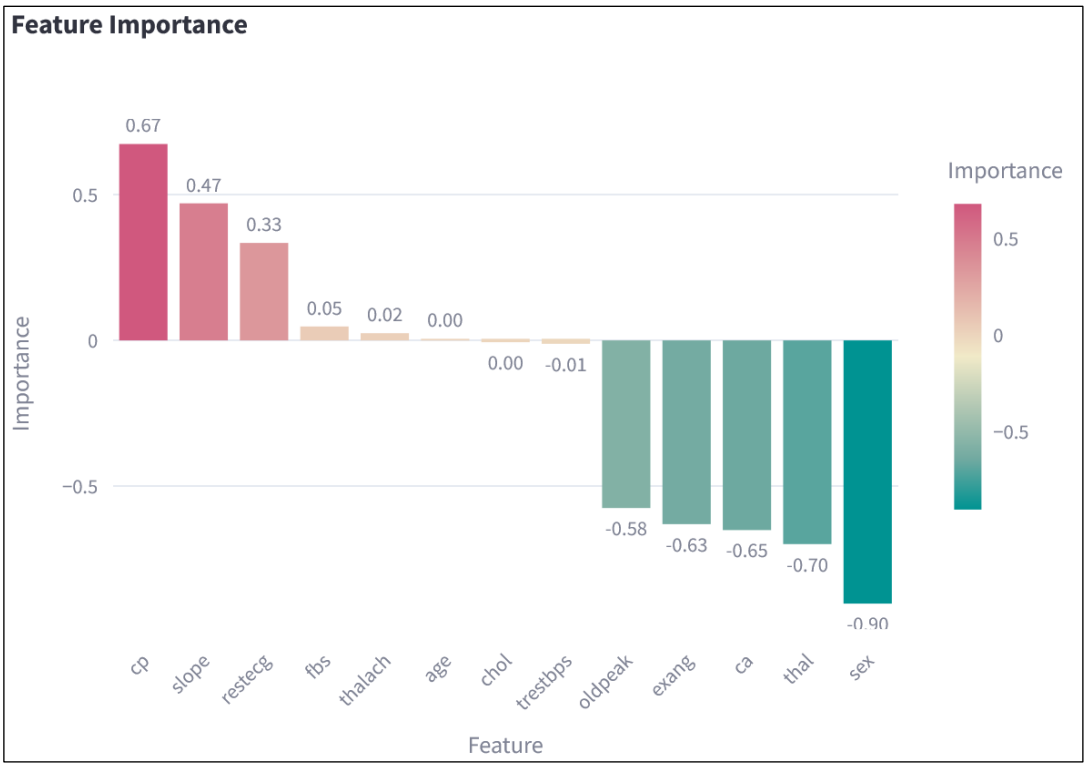

Рисунок 4.20– Важливість ознак для логістичної регресії (рисунок виконано самостійно)

Даний графік відображує ступінь, до якої кожна ознака впливає на процес прийняття рішень моделі під час визначення того, чи певні закономірності у вибірці даних про здоров'я пацієнта вказують на підвищену ймовірність серцевих захворювань.

Під час розгляду коефіцієнтів ознак ми помічаємо, що деякі з них негативні, а деякі позитивні. Більше значення вказує на те, що ознака має сильніший вплив на процес прийняття рішень у моделі. Від'ємні значення вказують на негативну кореляцію, тоді як позитивні значення вказують на позитивну кореляцію.

Наприклад, коефіцієнт для атрибута «стать» становить -0,904. Це від'ємне значення свідчить про те, що зі збільшенням значення «статі»(0 – жінка, 1 – чоловік) ймовірність досягнення цільового значення(target, де 0 – зорова людина, 1 – хвора) зменшується.

Ми можемо далі дослідити цей зв'язок, порівнявши стовпець «стать» зі стовпцем цільового значення (див. рис. 4.21).

target	0	1
sex		
0	24	72
1	114	93

Рисунок 4.21 – Таблиця порівняння статі і захворюваності (рисунок виконано самостійно)

При детальнішому розгляді ми спостерігаємо, що коли атрибут «стать» дорівнює 0 (що вказує на жінку), людей із захворюваннями серця приблизно втричі більше (ціль = 1) порівняно з тими, хто їх не має (72 проти 24).

І навпаки, коли атрибут «стать» збільшується до 1 (що вказує на чоловіка), співвідношення між людьми із захворюваннями серця та без них звужується майже до один до одного (114 проти 93).

Це вказує на те, що модель виявила помітну закономірність, що відображає набір даних. На основі цих спостережень та конкретного набору даних можна припустити, що жінки більш схильні до захворювань серця.

Давайте дослідимо наявність позитивної кореляції. Для цього створимо ще одну таблицю (див. рис. 4.22).

target	0	1
slope		
0	12	9
1	91	49
2	35	107

Рисунок 4.21 – Таблиця порівняння slope і захворюваності (рисунок виконано самостійно)

«Slope» представляє нахил сегмента ST пікового фізичного навантаження з трьома категоріями:

– 0: підвищений (краща частота серцевих скорочень під час фізичних вправ) – незвичайно;
– 1: плоский нахил (мінімальна зміна) - типово для здорового серця;
– 2: низхідний нахил (ознаки нездорового серця).

Згідно з нашою моделлю, існує позитивна кореляція 0,470 між «slope» та цільовою змінною, хоча й не така сильна, як кореляція між «статтю» та цільовою змінною, вона все ж заслуговує на увагу.

Ця позитивна кореляція свідчить про те, що зі збільшенням «slope» зростає і цільове значення.

ВИСНОВКИ

У межах даної роботи було досліджено та реалізовано підхід до створення системи автоматизованого прогнозування серцевого нападу за допомогою методів машинного навчання. Метою дослідження стало підвищення ефективності раннього виявлення ризиків серцево-судинних подій, що є критично важливим для своєчасного медичного втручання та зниження рівня смертності.

Аналіз сучасних підходів у галузі виявив значний потенціал моделей машинного навчання, таких як логістична регресія, метод k найближчих сусідів (KNN), випадковий ліс (Random Forest), у задачах медичної діагностики. Було реалізовано програмну систему, яка дозволяє здійснювати автоматичне навчання та підбір гіперпараметрів моделей з використанням GridSearchCV та RandomizedSearchCV, а також аналіз точності, повноти, F1-міри та площі під ROC-кривою (AUC) тощо.

Проведено практичне дослідження із використанням відкритого медичного набору даних (наприклад, Heart Disease Dataset), яке показало, що моделі машинного навчання здатні демонструвати високу точність у виявленні пацієнтів із ризиком серцевого нападу. Зокрема, модель логістичної регресії досягла найкращого балансу між точністю, чутливістю та специфічністю, що робить її придатною для медичного використання.

У ході реалізації були визначені ключові етапи:

- попередня обробка та візуалізація даних для аналізу особливостей медичних параметрів;
- побудова та навчання моделей з використанням різних алгоритмів;
- багатокритеріальний аналіз моделей за метриками точності, часу навчання та складності реалізації;
- реалізація інтерфейсу користувача для зручного використання моделі у практичних умовах.

Окрему увагу було приділено аналізу важливих ознак для передбачення серцевого нападу. Використання моделей з можливістю оцінки важливості ознак дозволило виявити, які ознаки впливають найбільше на прийняття рішень.

Серед основних викликів було визначено:

- залежність результатів від якості та повноти даних;
- потребу в балансуванні класів через дисбаланс позитивних та негативних випадків у датасеті;
- ризики перенавчання при роботі з обмеженим обсягом даних.

Підсумковий багатокритеріальний аналіз, проведений із використанням методу вагових коефіцієнтів, дозволив порівняти моделі за кількома критеріями. На основі в результаті було визначено, що модель логістичної регресії є найбільш ефективною з точки зору збалансованості якості передбачення та практичної придатності.

Загалом, результати дослідження підтвердили доцільність і перспективність використання методів машинного навчання для прогнозування серцевих нападів. Створена система може стати основою для впровадження інтелектуальних інструментів підтримки лікарських рішень у клінічній практиці, підвищуючи ефективність скринінгу та профілактики серцево-судинних захворювань.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Robin Genuer Jean-Michel Poggi “Random Forests with R”. URL: https://dokumen.pub/random-forests-with-r-9783030564841.html (дата звернення 10.06.2025).

2. Wade C. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing, 2020. 310 p.

3. Wyk A. V. Practical Machine Learning with LightGBM and Python: Explore Microsofts Gradient Boosting Framework to Optimize Machine Learning. Packt Publishing, 2023. 252 p.

4. Teslenko D. , Smelyakov K., “Role and evolution of computer vision in medicine”. URL: https://doi.org/10.36074/grail-of-science.15.03.2024.030 (дата звернення 10.06.2025).

5. Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. URL: https://link.springer.com/chapter/10.1007/978-981-16-3071-2_8 (дата звернення 10.06.2025).

6. Hasan O. S., Saleh I. A. Development of Heart Attack Prediction Model Based on Machine Learning. Eastern-European Journal of Enterprise Technologies. 2021. Vol. 112, no. 4/2. P. 26. URL: https://doi.org/10.15587/1729-4061.2021.238528 (дата звернення 10.06.2025).

7. Laftah R. H., Al-Saedi K. H. K. Explainable Ensemble Learning Models for Early Detection of Heart Disease. Journal of Robotics and Control. 2024. Vol. 5, no. 5. URL: https://doi.org/10.18196/jrc.v5i5.22448 (дата звернення 10.06.2025).

8. ECG Based Early Heart Attack Prediction Using Neural Networks / K. Ashish et al. IEEE, 2022. URL: https://doi.org/10.1109/ICESC54411.2022.9885448 (дата звернення 10.06.2025).

9. Mansourifar H., Weidong S. Deep Synthetic Minority Over-Sampling Technique. 2020. URL: https://doi.org/10.48550/arXiv.2003.09788 (дата звернення 10.06.2025).

10. Gruzdo I., Kyrychenko I., Tereshchenko G., Shanidze O., “Analysis of Models Usability Methods Used on Design Stage to Increase Site Optimization”. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163053422&partnerID=40&md5=46e44e3459e277732495dafe99c2e552 (дата звернення 10.06.2025).

11. GitHub – Heart Attack Prediction. GitHub. URL: https://github.com/OleksiiSandin/heart-attack-prediction (дата звернення 10.06.2025).

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

4. Teslenko D. , Smelyakov K., “Role and evolution of computer vision in medicine”. URL: https://doi.org/10.36074/grail-of-science.15.03.2024.030 (дата звернення 30.12.2024).

10. Gruzdo I., Kyrychenko I., Tereshchenko G., Shanidze O., “Analysis of Models Usability Methods Used on Design Stage to Increase Site Optimization”. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163053422&partnerID=40&md5=46e44e3459e277732495dafe99c2e552 (дата звернення 05.06.2025).