МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно- обчислювальних систем і управління

ЗАКАЛЮК Павло Андрійович

Прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання/Heart disease forecasting using ensemble machine learning methods

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп’ютерні науки

Кваліфікаційна робота

Виконав студент групи КН-41
П. А. Закалюк

Науковий керівник:
Ю. В. Константкевич

Кваліфікаційну роботу
допущено до захисту:

"___" _____ 20___ р.

Завідувач кафедри
М. П. Комар

ТЕРНОПІЛЬ - 2024

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «бакалавр»
спеціальність 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
ЗАКАЛЮКУ Павлу Андрійовичу
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи: Прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання / Heart disease forecasting using ensemble machine learning methods
керівник роботи Констанкевич Юрій Володимирович, викладач
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від 12 грудня 2023 р. № 753.
2. Строк подання студентом закінченої кваліфікаційної роботи 15 травня 2024 р.
3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.
4. Основні питання, які потрібно розробити:
– Провести детальний аналіз сучасних методів діагностики серцевих захворювань, оцінити їхні переваги та недоліки.
– Дослідити можливості застосування машинного навчання в медицині, зокрема в контексті серцево-судинних захворювань.
– Проаналізувати існуючі рішення для прогнозування серцевих захворювань за допомогою машинного навчання і визначити їх обмеження.
– Обрати оптимальні ансамблеві методи машинного навчання для підвищення точності прогнозування серцевих захворювань.
– Розробити архітектуру модуля прогнозування серцевих захворювань та визначити математичні моделі для ансамблевих методів.
– Провести порівняльний аналіз різних ансамблевих методів на реальних медичних даних, включаючи оцінку ефективності кожного методу.
– Реалізувати модуль прогнозування серцевих захворювань і перевірити її дієвість, порівнявши результати з традиційними методами.

5. Перелік графічного матеріалу в роботі:
– Архітектура модуля прогнозування серцевих захворювань
– Алгоритм прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання
– графіки з результатами експериментальних досліджень.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 12 грудня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2024 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2024 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 01.04.2024 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 01.05. 2024 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 15.05.2024 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 20.05.2024 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту у системі «Unicheck».	до 10.06.2024 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 14.06.2024 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.06. 2024 р.	

Студент_____ П. А. Закалюк
(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи_____ Ю.В. Констанкевич
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Кваліфікаційна робота на тему «Прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання» на здобуття освітнього ступеня «бакалавр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 40 сторінок і містить 19 ілюстрацій, 2 таблиці, 2 додатки та 12 використаних джерел.

Метою роботи є розробка ефективної моделі прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання.

Методами розроблення обрано метод аналізу (для дослідження існуючих підходів до прогнозування), метод синтезу (для поєднання переваг існуючих методів), методи моделювання (для представлення та дослідження процесів прогнозування), метод порівняльного аналізу (для оцінювання адекватності моделі прогнозування).

Внаслідок виконання роботи обґрунтовано раціональний підхід до розроблення моделей прогнозування серцевих захворювань та розроблено програмний засіб, який дозволяє створювати і досліджувати моделі прогнозування серцевих захворювань.

Результати дослідження можуть бути використані в науково-дослідних установах і підрозділах підприємств, що займаються розробленням моделей прогнозування.

Ключові слова: СЕРЦЕВІ ЗАХВОРЮВАННЯ, МАШИННЕ НАВЧАННЯ, АНСАМБЛЕВІ МЕТОДИ, ПРОГНОЗУВАННЯ.

ANNOTATION

Qualification work on the topic «Heart disease forecasting using ensemble machine learning methods» for Bachelor's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 40 pages and it contains 19 figures, 2 tables, 2 annexes and 12 sources.

The purpose of the work is to develop an effective model for forecasting heart diseases using ensemble machine learning methods.

Research methods include analysis (to study existing approaches to forecasting), synthesis (to combine the advantages of existing methods), modeling (to represent and study the forecasting processes), and comparative analysis (to evaluate the adequacy of the forecasting model).

As a result of the work, a rational approach to the development of heart disease forecasting models was substantiated, and a software tool was developed that allows creating and researching heart disease forecasting models.

The research results can be used in research institutions and enterprise departments involved in the development of forecasting models.

Keywords: HEART DISEASES, MACHINE LEARNING, ENSEMBLE METHODS, FORECASTING.

ЗМІСТ

ВСТУП

Актуальність дослідження прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання обумовлена кількома ключовими факторами. По-перше, серцево-судинні захворювання залишаються однією з головних причин смертності та інвалідності у світі, що зумовлює потребу у їх ранньому виявленні та ефективному лікуванні. Розвиток та впровадження новітніх технологій у медичну практику може значно покращити якість та доступність медичних послуг, зокрема у профілактиці та діагностиці.

По-друге, традиційні методи діагностики часто вимагають значних часових та фінансових витрат, а також залежать від суб'єктивної оцінки медичних фахівців, що може призводити до помилок. Ансамблеві методи машинного навчання, які об'єднують декілька алгоритмів для досягнення більшої точності та надійності прогнозу, виявляються особливо ефективними в автоматизації аналізу медичних даних. Використання цих методів дозволяє знижувати ризики людської помилки та забезпечує більш точне та об'єктивне оцінювання стану здоров'я пацієнтів.

По-третє, інтеграція машинного навчання в медичні інформаційні системи сприяє розвитку персоналізованої медицини. Ансамблеві методи можуть адаптуватися до індивідуальних особливостей пацієнтів, що дозволяє підвищити ефективність лікувальних інтервенцій і запобігти розвитку ускладнень. Таким чином, дослідження в цій галузі відкриває нові перспективи для покращення якості життя та здоров'я населення, роблячи медичну допомогу більш точною, доступною та ефективною.

Отже, метою бакалаврської роботи є оцінка ефективності ансамблевих методів машинного навчання в прогнозуванні серцевих захворювань для покращення діагностичних можливостей та лікувальних стратегій.

Для досягнення мети бакалаврської роботи необхідно виконати наступні завдання:

1. Провести детальний аналіз сучасних методів діагностики серцевих захворювань, оцінити їхні переваги та недоліки.
2. Дослідити можливості застосування машинного навчання в медицині, зокрема в контексті серцево-судинних захворювань.
3. Проаналізувати існуючі рішення для прогнозування серцевих захворювань за допомогою машинного навчання і визначити їх обмеження.
4. Обрати оптимальні ансамблеві методи машинного навчання для підвищення точності прогнозування серцевих захворювань.
5. Розробити архітектуру модуля прогнозування серцевих захворювань та визначити математичні моделі для ансамблевих методів.
6. Провести порівняльний аналіз різних ансамблевих методів на реальних медичних даних, включаючи оцінку ефективності кожного методу.
7. Реалізувати модуль прогнозування серцевих захворювань і перевірити її дієвість, порівнявши результати з традиційними методами.

Об'єктом дослідження є процес прогнозування серцевих захворювань на основі аналізу медичних даних за допомогою машинного навчання.

Предметом дослідження є методи ансамблевого машинного навчання, застосовані для підвищення точності та ефективності діагностики серцевих захворювань.

Методи дослідження в даній роботі охоплюють широкий спектр технік машинного навчання, зокрема ансамблеві методи такі як бустинг, беггінг та стекінг. Для аналізу та обробки даних застосовуються статистичні методи, обчислювальні алгоритми та програмне забезпечення для обробки великих даних. Окрім того, використовуються методи валідації моделей, такі як перехресна перевірка, для оцінки їх точності та надійності. Дослідження також включає методи математичного моделювання для формулювання та

оптимізації ансамблевих моделей, що сприяє глибокому розумінню механізмів захворювань і взаємодії між біомаркерами.

Практичне значення даної бакалаврської роботи полягає в покращенні процесів діагностики та прогнозування серцевих захворювань, що може забезпечити більш раннє виявлення потенційних патологій та оптимізувати підходи до лікування. Впровадження розроблених ансамблевих методів в клінічну практику може допомогти медичним фахівцям ухвалювати обґрунтованіші та ефективніші клінічні рішення, знижувати витрати на лікування за рахунок оптимізації терапевтичних втручань і, відповідно, підвищувати якість життя пацієнтів. Це також може вплинути на подальші дослідження в області медицини та машинного навчання, надаючи основу для нових інноваційних розробок.

Структура та обсяг кваліфікаційної роботи. Робота складається зі вступу, трьох розділів, висновків і списку використаних джерел. Загальний обсяг становить 40 сторінок комп'ютерного тексту, включаючи 19 рисунків і 2 таблиці. Список використаних джерел містить 12 найменувань і займає 2 сторінки.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися V Всеукраїнської мультидисциплінарної студентської наукової конференції «Науковий простір: аналіз, сучасний стан, тренди та перспективи», яка відбулася 17 травня 2024 року у місті Київ, Україна.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Сучасні підходи в діагностиці серцевих захворювань

Огляд предметної області серцевих захворювань підкреслює важливість розуміння серцево-судинної системи у контексті глобального здоров'я. Серцеві захворювання залишаються однією з провідних причин смерті у світі, що вимагає комплексного підходу до їх виявлення та лікування. Розуміння механізмів розвитку таких станів, як ішемічна хвороба серця, серцева недостатність та аритмії, є ключовим для розробки ефективних стратегій профілактики та інтервенцій.

Інновації в медичній науці та технологіях значно підвищують якість догляду за пацієнтами з серцевими захворюваннями. Використання передових технологій в медицині дозволяє клініцям забезпечувати більш індивідуалізовані та ефективні лікувальні плани, що спираються на детальний аналіз здоров'я пацієнта. Такий підхід допомагає знизити ризики та покращує результати лікування, включаючи зниження ймовірності рецидивів та подальших ускладнень.

Збільшення обізнаності про ризики та симптоми серцевих захворювань також відіграє важливу роль у підвищенні рівня загального здоров'я населення. Просвітницькі програми та кампанії здорового способу життя сприяють ранньому виявленню та профілактиці, зменшуючи кількість випадків важких захворювань та надзвичайних станів. Освітні ініціативи можуть включати інформацію про важливість регулярних медичних оглядів, користь фізичної активності та здорового харчування.

Наукові дослідження в галузі кардіології неухильно рухаються вперед, зосереджуючись на виявленні нових біомаркерів та генетичних факторів, які можуть сприяти розвитку серцевих захворювань. Ці знахідки відкривають

можливості для розробки нових лікарських засобів та терапевтичних стратегій, що цілеспрямовано впливають на молекулярні механізми захворювань. Такий напрямок досліджень може призвести до створення персоналізованих лікувань, які будуть ефективнішими для конкретних пацієнтів.

Співпраця між кардіологами, дослідниками, охороною здоров'я та пацієнтами є фундаментальною для успіху в боротьбі з серцевими захворюваннями. Мультидисциплінарний підхід дозволяє об'єднувати зусилля в навчанні, дослідженнях та практичному застосуванні медичних знань, що в свою чергу сприяє кращому розумінню та управлінню серцево-судинними захворюваннями на всіх рівнях.

Сучасні підходи (Рисунок 1.1) до діагностики серцевих захворювань включають широкий спектр технологій та методів, що дозволяють лікарям точно виявляти та моніторити стани серцево-судинної системи. Основні технології та методи, які активно використовуються в клінічній практиці, охоплюють наступне:

1. Електрокардіографія (ЕКГ). Цей метод є стандартною процедурою при оцінці серцевої активності, який вимірює електричну активність серця через електроди, розміщені на шкірі пацієнта. ЕКГ допомагає виявити аритмії, ішемію та інші порушення, які можуть свідчити про ризик розвитку серцевих захворювань.

2. Ехокардіографія. Цей ультразвуковий метод дозволяє візуалізувати структуру і функцію серця. Використання звукових хвиль допомагає лікарям оцінити розмір камер серця, стан серцевих клапанів, а також наявність будь-яких утворень чи вад серця.

3. Комп’ютерна томографія (КТ) серця КТ серця використовується для більш детального вивчення коронарних артерій та оцінки ступеню коронарного атеросклерозу. Цей метод є особливо корисним для виявлення звужень артерій, що може призводити до серцевих нападів.

4. Магнітно-резонансна томографія (МРТ) серця МРТ серця забезпечує високу якість візуалізації серцевого м’яза, включно з визначенням областей ішемії та фіброзу. Метод допомагає оцінити не тільки структуру, але і функціональні аспекти серця, такі як викидний фракцій.

5. Напрямні посилених даних (Wearable Technology) Останнім часом зросла популярність використання портативних пристроїв, що можуть постійно моніторити життєві показники, такі як частота серцевих скорочень та ритм. Ці дані можуть використовуватися для раннього виявлення потенційних проблем, навіть коли пацієнт не знаходиться в лікарні.

6. Лабораторні біомаркери Вимірювання рівнів різних біомаркерів у крові, таких як тропоніни, натріюретичні пептиди та інші, може допомогати в діагностиці та моніторингу серцевої недостатності та інфаркту міокарда.

7. Машинне навчання і штучний інтелект Сучасні дослідження інтегрують машинне навчання для аналізу великих даних, отриманих з різних діагностичних методів, щоб передбачити ризик розвитку серцевих захворювань або оцінити реакцію на лікування. Це включає аналіз медичних записів, образів, ЕКГ та інших біомаркерів для створення точніших моделей прогнозування.

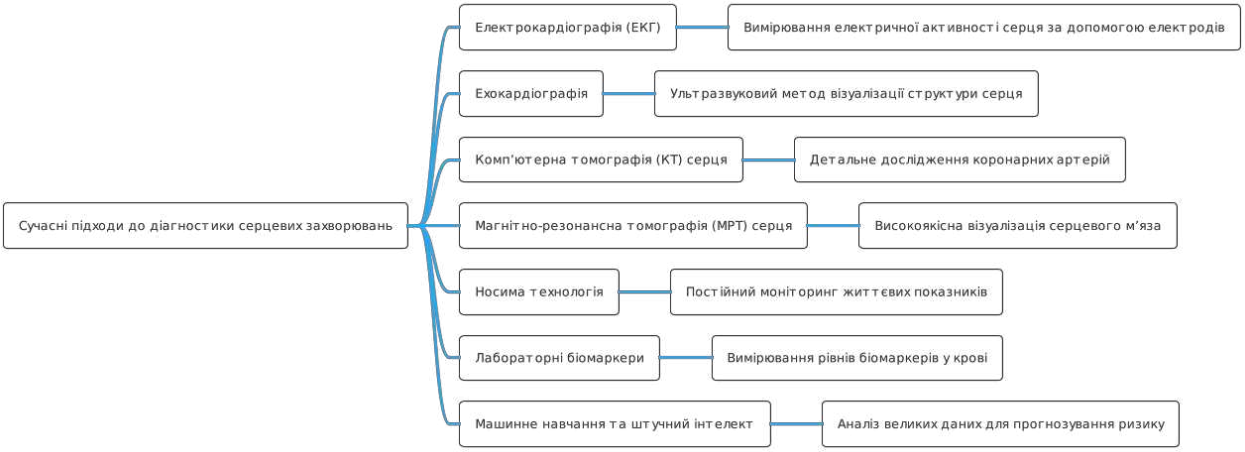


Рисунок 1.1 - Сучасні підходи до діагностики серцевих захворювань

Сучасні підходи до діагностики серцевих захворювань використовують широкий спектр технологій та методів, що забезпечують комплексне та точне вивчення стану серцево-судинної системи. Від традиційних методів, таких як електрокардіографія (ЕКГ) та ехокардіографія, до передових технологій, як-от комп’ютерна томографія (КТ) та магнітно-резонансна томографія (МРТ) серця, кожен з цих методів має свої переваги у виявленні різних аспектів серцевої діяльності. Використання напрямних посилених даних, таких як портативні пристрої для моніторингу, разом із лабораторними біомаркерами та інтеграцією машинного навчання і штучного інтелекту, збільшує потенціал раннього виявлення серцевих захворювань та їх профілактики. Ці технології разом формують новітню еру в кардіології, що дозволяє медичним фахівцям не тільки діагностувати, але й ефективно планувати лікування, враховуючи індивідуальні особливості кожного пацієнта.

1.2 Застосування машинного навчання в медицині

Застосування машинного навчання в медицині відкриває нові можливості для покращення діагностики, лікування та профілактики захворювань. Особливо це стосується галузі кардіології, де аналіз великих обсягів даних може допомогти в ранньому виявленні патологій серцево-судинної системи та їх ефективному лікуванні. Ось детальний огляд ключових аспектів застосування машинного навчання в медицині (Рисунок 1.2):

1. Аналіз медичних даних. Машинне навчання дозволяє обробляти та аналізувати великі набори медичних даних, включаючи медичні записи, результати лабораторних аналізів, зображення, отримані за допомогою різноманітних діагностичних інструментів (наприклад, МРТ, КТ, ультразвук). Алгоритми машинного навчання можуть ідентифікувати закономірності та зв'язки, які не завжди очевидні для людського ока чи традиційних аналітичних методів.

2. Підтримка прийняття рішень. Штучний інтелект може служити потужним інструментом підтримки прийняття рішень для лікарів, зокрема в складних або критичних ситуаціях. Системи, засновані на машинному навчанні, можуть пропонувати рекомендації щодо лікування, базуючись на історії хвороби, генетичних данних та інших важливих параметрах. Це допомагає лікарям вибирати оптимальний підхід до лікування конкретного пацієнта.

Рисунок 1.2 - Застосування машинного навчання в медицині

3. Прогнозування розвитку захворювань. Алгоритми машинного навчання використовуються для розробки моделей, які можуть прогнозувати ризик розвитку певних захворювань, засновуючись на різноманітних біомаркерах та інших факторах ризику. Наприклад, в кардіології це може стосуватися ризику інфаркту міокарда або інсульту. Ці моделі допомагають в ранньому виявленні пацієнтів з високим ризиком та можуть запропонувати втручання на ранніх стадіях.

4. Персоналізована медицина. Машинне навчання є ключем до розвитку персоналізованої медицини, де лікування та профілактичні заходи можуть бути адаптовані під конкретні генетичні особливості та потреби пацієнта. Це

особливо важливо у випадках лікування хронічних захворювань або умов, що вимагають довгострокового управління, наприклад, діабет або гіпертонія.

5. Автоматизація рутинних процесів. Машинне навчання дозволяє автоматизувати багато рутинних та часомістких медичних процесів, таких як введення даних, обробка медичних зображень та управління медичними записами. Це звільняє медичних працівників від рутинних задач і дозволяє їм зосередитись на більш важливих аспектах пацієнтського догляду.

Застосування машинного навчання в медицині продовжує розвиватись, і майбутнє цієї технології обіцяє ще більші можливості для покращення здоров'я та благополуччя пацієнтів по всьому світу.

Машинне навчання в медицині використовує різноманітні методи, кожен з яких має свої переваги та недоліки в контексті певних задач. Однак, ансамблеві методи часто виокремлюються як найбільш ефективні, особливо коли мова йде про забезпечення високої точності та надійності в прогнозуванні. Ось детальний огляд (Рисунок 1.3) основних методів машинного навчання, що застосовуються в медицині:

Рисунок 1.3 - Методи машинного навчання в медицині

1. Навчання з вчителем. Цей метод використовується для класифікації пацієнтів або прогнозування результатів лікування на основі попередньо розмічених даних. Наприклад, алгоритми наглядового навчання можуть аналізувати медичні зображення, такі як МРТ, для виявлення пухлин або інших патологій.

2. Навчання без вчителя. Цей підхід використовується для виявлення закономірностей або кластерів у великих наборах даних без попереднього визначення міток. У медицині це може бути корисно для ідентифікації підгруп пацієнтів з подібними симптомами або реакціями на лікування.

3. Напівнаглядове навчання. Комбінує наглядове і ненаглядове навчання для використання як розмічених, так і нерозмічених даних. Це може покращити точність моделей у випадках, коли розмічені дані обмежені або дорогі в отриманні.

4. Посилювальне навчання. Використовується для розробки стратегій рішення на основі послідовності рішень, які ведуть до найкращого результату. Це може бути застосовано для оптимізації лікувальних процедур або для автоматизації роботи медичних роботів.

Ансамблеві методи, такі як Random Forests, Gradient Boosting Machines (GBM) та Stacking, використовують кілька навчальних алгоритмів для досягнення кращого прогнозування. Вони ефективно комбінують прогнози з множини моделей, зменшуючи помилки та підвищуючи загальну точність.

Ансамблеві методи вважаються одними з найефективніших у машинному навчанні, особливо у медичних застосуваннях, де точність є критично важливою. Вони дозволяють знизити вплив випадкових помилок одиничних моделей та забезпечують більш стабільні та надійні результати. Особливо ефективними вони є у випадках, коли потрібно обробляти складні набори даних із широким спектром змінних. Таким чином, ансамблевий підхід може вважатися найкращим вибором для розробки високоефективних медичних систем на основі машинного навчання.

1.3 Огляд існуючих рішень

У сучасному світі, де серцево-судинні захворювання залишаються однією з головних причин смертності глобально, важливість точного та своєчасного діагностування не може бути недооцінена. Застосування машинного навчання у медичній сфері відкриває нові можливості для раннього виявлення та кращого розуміння хвороб серця. Для оцінки поточного стану досліджень в цій області, нами було складено порівняльну таблицю, що включає різноманітні підходи та методології від провідних науковців у галузі.

Дослідження Prasad [1] та співавторів розглядає новий підхід до прогнозування серцевих нападів за допомогою генетичних алгоритмів та методів оптимізації зграї частинок (PSO) у комбінації з машинним навчанням. Ця робота використовує процес відбору ознак для точного передбачення серцевих захворювань.

Робота [2] N Bhattacharya та P Kumar досліджує використання аналітики охорони здоров’я для покращення управління діабетом та запобігання серцевих нападів. Вони застосовують машинне навчання та системи підтримки клінічних рішень для аналізу великих баз даних і передбачення серцевих захворювань.

Стаття [3] MA Islam та співавторів зосереджується на застосуванні алгоритмів машинного навчання та стратегій відбору ознак для точного прогнозування серцевих захворювань. Вони вносять вклад у зменшення захворюваності, пов’язаної з серцевими хворобами.

Дослідження [4] EJ Topol описує використання штучного інтелекту для інтерпретації медичних знімків, які точно прогнозують різні типи та тяжкість клапанних серцевих захворювань за допомогою тренування моделей на базі понад 22 000 спарених ехокардіограм.

Робота [5] B Booba вивчає застосування гібридних ансамблевих моделей машинного навчання для аналізу діабету з метою раннього виявлення та запобігання серцевих захворювань. У дослідженні використовуються декілька алгоритмів машинного навчання для точного прогнозування.

Стаття [6] N Bhattacharya та P Kumar досліджує інтеграцію аналітики охорони здоров'я для покращення управління діабетом та запобігання серцевих нападів. Дослідження використовує методи видобування даних, машинне навчання та системи підтримки клінічних рішень для аналізу великих баз даних і передбачення захворювань серця.

Дослідження [7] MA Islam та колег фокусується на глибокому аналізі алгоритмів машинного навчання та стратегій відбору ознак для точного прогнозування серцевих захворювань. Дослідження показує, що використання відбору ознак покращує здатність моделей до виявлення значущих патернів, що стосуються діагностики серцевих захворювань.

Стаття [8] M Wang та співавторів розглядає використання моделі глибокого навчання MTU-Net3+ для автоматизованого аналізу базових/прискорювальних/сповільнювальних характеристик частоти серцевих скорочень плоду (FHR). Модель призначена для багатозадачного аналізу FHR, що сприяє точнішій діагностиці станів плоду.

Дослідження [9] P Kaushal та S Singh описує використання трансформера на основі енкодера для аналізу виживаності на прикладі даних про серцеву недостатність з правостороннім цензуруванням. Це дослідження акцентує увагу на важливості моделей аналізу виживаності у передбаченні серцевої недостатності.

Стаття [10] RJH Miller та інші зосереджується на використанні глибокого навчання для кількісного аналізу 99mTc-пірофосфатної SPECT/CT для діагностики кардіоамілоїдозу. В рамках вторинного аналізу оцінюються асоціації з комбінованим клінічним результатом серцево-судинної смерті або госпіталізації через серцеву недостатність.

Таблиця 1.1 демонструє широкий спектр досліджень, кожне з яких приносить унікальний вклад у розвиток діагностики серцевих захворювань. Від генетичних алгоритмів та оптимізації зграї частинок (PSO) у роботі Prasad et al. [1], до інтеграції штучного інтелекту для інтерпретації медичних знімків у дослідженні EJ Topol [4], кожен підхід має свої переваги та обмеження. Наше власне дослідження, що також зосереджене на машинному навчанні для прогнозування серцевих захворювань, спирається на ці знахідки, використовуючи передові техніки для покращення точності прогнозів.

Таблиця 1.1 - Порівняльний огляд

Автори / Дослідження	**Фокус дослідження**	**Методологія**	**Основні висновки**
Prasad et al. [1]	Прогнозування серцевих нападів	Генетичні алгоритми та PSO	Використання процесу відбору ознак для точного передбачення серцевих захворювань
Bhattacharya & Kumar [2]	Покращення управління діабетом та запобігання серцевим нападам	Машинне навчання, системи підтримки клінічних рішень	Аналіз великих баз даних для передбачення серцевих захворювань
MA Islam et al. [3]	Прогнозування серцевих захворювань	Алгоритми машинного навчання, стратегії відбору ознак	Зменшення захворюваності на серцеві хвороби
EJ Topol [4]	Інтерпретація медичних знімків	Штучний інтелект, тренування моделей на базі ехокардіограм	Точне прогнозування різних типів та тяжкості клапанних серцевих захворювань
B Booba [5]	Аналіз діабету та серцевих захворювань	Гібридні ансамблеві моделі машинного навчання	Раннє виявлення та запобігання серцевих захворювань за допомогою кількох алгоритмів машинного навчання

Незважаючи на різноманітність у підходах, усі дослідження прагнуть зменшити захворюваність і смертність від серцевих захворювань за допомогою інноваційних технологічних рішень. Наше дослідження вносить вклад у це поле, надаючи додаткові дані та результати, які можуть бути використані для подальшого розвитку цих технологій. Така інтеграція різноманітних досліджень дозволяє створити комплексний підхід до лікування

та профілактики серцевих захворювань, що є критично важливим для подальшого зниження рівня смертності на глобальному рівні.

1.4 Вибір перспективного шляху і постановка задачі дослідження

Прогнозування серцевих захворювань за допомогою ансамблевих методів машинного навчання відіграє ключову роль у сучасній медичній практиці. Актуальність такого дослідження полягає у високій поширеності серцево-судинних захворювань, які є однією з основних причин смертності на глобальному рівні. Впровадження передових технологій у медичну діагностику дозволяє не тільки підвищувати точність діагностики, але й значно покращувати результати лікування пацієнтів завдяки своєчасному виявленню ризиків.

Застосування ансамблевих методів, таких як випадкові ліси, бустинг або стекінг, в розробці прогностичних моделей, виявляється особливо ефективним у медичних застосуваннях. Ці методи базуються на комбінації декількох алгоритмів машинного навчання, що забезпечує високу точність прогнозів завдяки редукції варіативності та помилок, характерних для окремих моделей. У контексті серцевих захворювань це дозволяє не тільки виявляти потенційні ризики, але й адаптувати стратегії лікування до індивідуальних особливостей пацієнтів.

Окрім підвищення точності, ансамблеві методи також сприяють покращенню загальної роботи системи здоров’я шляхом інтеграції більш динамічних та гнучких моделей у клінічну практику. Вони дозволяють медичним фахівцям оперативно реагувати на зміни стану пацієнта, що є особливо важливим у лікуванні хронічних умов, таких як серцева недостатність або атеросклероз.

Наукове співтовариство активно розглядає питання подальшого розвитку та оптимізації ансамблевих методів, а також їх інтеграції з іншими

технологіями, наприклад, геномними даними, що може призвести до новаторських підходів у персоналізованій медицині. Особливий інтерес представляє розробка таких систем, які б могли автоматично адаптуватися до нових даних і вдосконалювати свої прогнози в реальному часі.

З урахуванням усіх цих аспектів, наше дослідження покликане не тільки дослідити потенціал ансамблевих методів у діагностиці серцевих захворювань, але й встановити основу для майбутніх інновацій у медичній галузі, що сприятиме покращенню якості життя пацієнтів на глобальному рівні.

Метою цієї бакалаврської роботи є оцінка ефективності ансамблевих методів машинного навчання в прогнозуванні серцевих захворювань для покращення діагностичних можливостей та лікувальних стратегій.

Для досягнення мети бакалаврської роботи необхідно виконати наступні завдання:

1. Провести детальний аналіз сучасних методів діагностики серцевих захворювань, оцінити їхні переваги та недоліки.
2. Дослідити можливості застосування машинного навчання в медицині, зокрема в контексті серцево-судинних захворювань.
3. Проаналізувати існуючі рішення для прогнозування серцевих захворювань за допомогою машинного навчання і визначити їх обмеження.
4. Обрати оптимальні ансамблеві методи машинного навчання для підвищення точності прогнозування серцевих захворювань.
5. Розробити архітектуру модуля прогнозування серцевих захворювань та визначити математичні моделі для ансамблевих методів.
6. Провести порівняльний аналіз різних ансамблевих методів на реальних медичних даних, включаючи оцінку ефективності кожного методу.

7. Реалізувати модуль прогнозування серцевих захворювань і перевірити її дієвість, порівнявши результати з традиційними методами.

Ці завдання створюють основу для дослідження, що сприяє розробці надійних і точних медичних діагностичних інструментів, здатних підвищити ефективність лікування та профілактики серцевих захворювань.

2 АЛГОРИТМІЧНЕ ЗАБЕЗПЕЧЕННЯ ПРОГНОЗУВАННЯ СЕРЦЕВИХ ЗАХВОРЮВАНЬ З ВИКОРИСТАННЯМ АНСАМБЛЕВИХ МЕТОДІВ МАШИННОГО НАВЧАННЯ

2.1 Архітектура модуля прогнозування серцевих захворювань

Прогнозування серцевих захворювань в сучасному світі викликає значний інтерес серед науковців та медичних фахівців, оскільки це дозволяє значно підвищити ефективність лікування та профілактики цих захворювань. Завдяки використанню передових технологій машинного навчання та збору великих обсягів медичних даних, можливо створити системи, здатні з високою точністю прогнозувати ризик розвитку серцевих захворювань у пацієнтів. Цей процес включає декілька критичних етапів, починаючи зі збору і перевірки даних, попередньої обробки, вибору відповідних ознак, тренування моделей і закінчуючи розгортанням готової моделі для використання в реальних умовах.

Прогнозування серцевих захворювань складається з декількох основних компонентів, які взаємодіють для забезпечення точного аналізу та прогнозування. Перший етап в системі - це збір даних, де здійснюється збір релевантної інформації, яка може включати клінічні дані, результати медичних обстежень, а також історію хвороб пацієнтів. Цей компонент також забезпечує перевірку зібраних даних на предмет їх повноти та достовірності, що є критично важливим для подальшого аналізу.

Наступний крок, попередня обробка даних, включає видалення неповних або аномальних записів, нормалізацію числових значень та розділення набору даних на тренувальні та тестові підмножини. Цей етап підготовки є фундаментальним для забезпечення того, що наступні моделі навчання машин будуть працювати з чистими та структурованими даними, що значно підвищує якість та надійність прогнозів.

Компонент вибір ознак відповідає за аналіз та вибір найбільш інформативних атрибутів з даних, які будуть використовуватися для

тренування моделей. Цей процес включає статистичний аналіз та застосування алгоритмів вибору ознак, щоб визначити, які ознаки мають найсильніший вплив на результати прогнозування. Вибір правильних ознак збільшує точність моделі та ефективність навчання.

Навчання моделі є центральним компонентом системи, де використовуються обрані ознаки для тренування алгоритмів машинного навчання. В рамках цього компонента відбувається оптимізація параметрів моделі, включаючи налаштування гіперпараметрів та вибір найефективнішої моделі на основі перехресної перевірки та інших методів валідації.

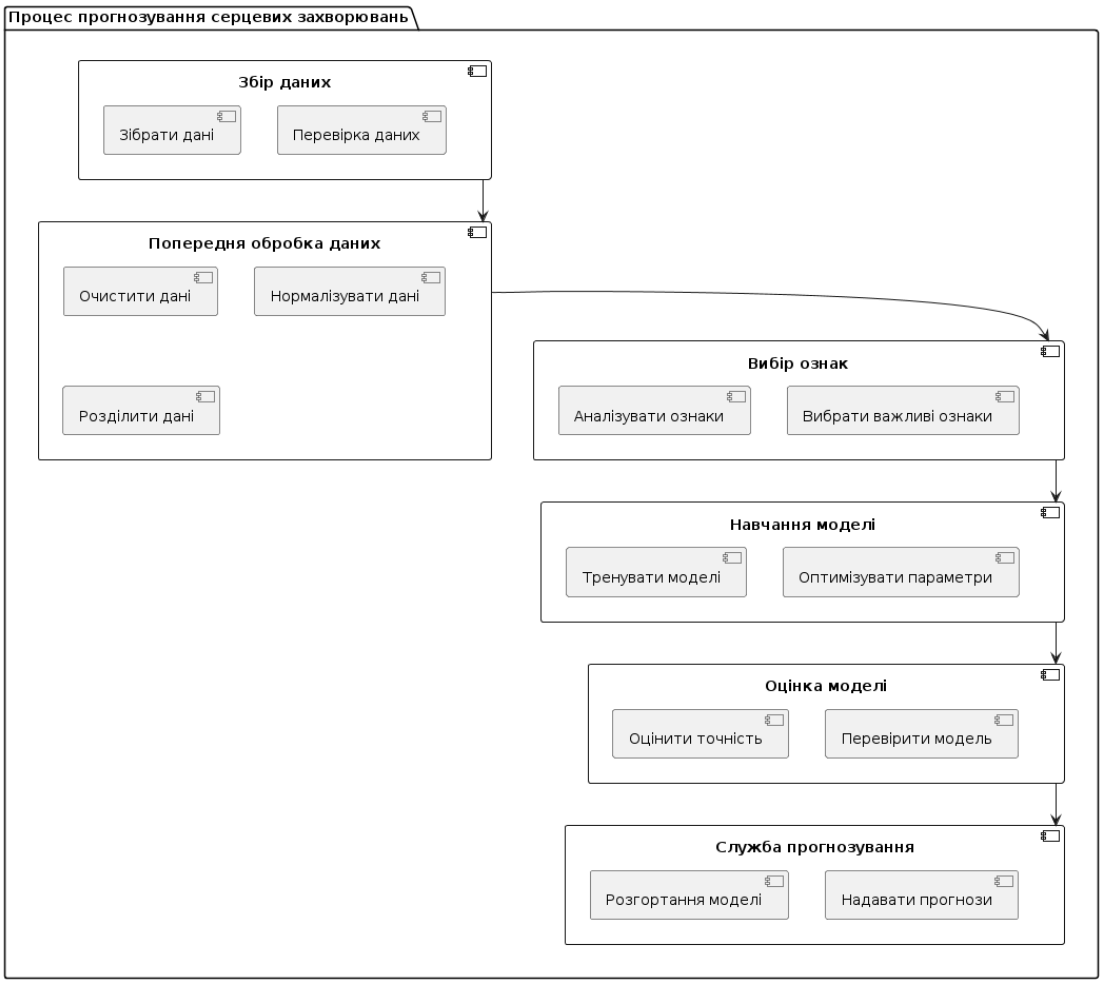


Рисунок 2.1 - Архітектура модуля прогнозування серцевих захворювань

Останній компонент, служба прогнозування, включає в себе розгортання тренованої моделі у виробниче середовище, де вона може надавати прогнози в реальному часі. Також цей компонент забезпечує моніторинг і оцінку

продуктивності моделі, що дозволяє своєчасно виявляти та коригувати будь-які відхилення у точності прогнозування.

Архітектура системи прогнозування серцевих захворювань, представлена на Рисунку 2.1, демонструє комплексний підхід до обробки та аналізу медичних даних для прогнозування. Кожен компонент системи відіграє ключову роль у забезпеченні точності та надійності прогнозу, де кінцевий компонент — служба прогнозування — забезпечує впровадження моделі в клінічну практику. Успішне використання такої системи може значно зменшити ризики для здоров'я пацієнтів, оптимізувати процеси медичного обслуговування та підвищити загальну якість життя популяції, що робить вивчення та розробку цих технологій надзвичайно важливою задачею сучасної медицини.

2.2 Математичне формулювання стекінг-ансамблю

Стекінг (англ. stacking, що ще називають стекованою узагальненням) — це ансамблевий метод машинного навчання, який поєднує прогнози з декількох моделей для формування більш точного кінцевого прогнозу за допомогою мета-класифікатора. Основна ідея стекінгу полягає у використанні прогнозів як вхідних даних для мета-класифікатора, який навчається визначати, як краще комбінувати ці прогнози для підвищення загальної точності. Нижче представлена математична модель стекінгу:

Нехай у нас є набір базових моделей $M = \{m_1, m_2, \dots, m_k\}$, де кожна модель *mi* робить прогноз $\hat{y}_i$ на основі навчального набору даних *X*.

Вихідні дані цих моделей, $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_k\}$, використовуються як вхідні дані (ознаки) для мета-класифікатора

Мета-класифікатор *m* meta навчається на основі $\hat{Y}$ і цільової змінної *Y*, щоб зробити кінцевий прогноз $\hat{y}$ meta . Функція втрат *L* мінімізується під час навчання мета-класифікатора:

$$L(Y, \hat{y}_{meta})$$

Для уникнення перенавчання, стекінг зазвичай використовує метод перехресної перевірки (cross-validation). На етапі тренування базових моделей використовують крос-валідацію для генерації прогнозів, які слугують вхідними даними для мета-класифікатора. Це дозволяє мета-класифікатору навчитися на різних взірцях прогнозів базових моделей, не перетинаючись з тренувальними даними, на яких ці базові моделі навчались.

Важливим аспектом є налаштування гіперпараметрів базових моделей та мета-класифікатора. Оптимальний набір параметрів може бути визначений через пошук по сітці або інші методи оптимізації.

Стекінг є ефективним засобом підвищення точності прогнозування, використовуючи силу кількох моделей та їх різноманітність, а мета-класифікатор слугує раціональним способом об'єднання цих прогнозів у вищий рівень точності.

2.3 Математична інтерпретація класифікаторів для стекінг-ансамблю

Стекінг-ансамбль використовує декілька базових класифікаторів для генерації прогнозів, які потім використовуються як вхідні дані для мета-класифікатора. Математично це можна представити так:

Базові класифікатори:

Нехай у нас є набір навчальних даних X, де кожен елемент $x_i \in X$ має мітку класу y_i. Базові класифікатори $\{m_1, m_2, \ldots, m_k\}$ використовуються для навчання на X та вироблення прогнозів $\hat{y}_i(j)$ для кожного класифікатора j.

1. XGBoost (xgb):

$$\hat{y}_i^{(xgb)} = f_{xgb}(x_i; \theta_{xgb})$$

де f_{xgb} - функція прогнозування XGBoost, параметризована параметрами θ_{xgb}, які мінімізують втрату через градієнтний бустінг.

2. k-Nearest Neighbors (knn):

$$\hat{y}_i^{(knn)} = \frac{1}{k} \sum_{x_j \in N_k(x_i)} y_j$$

де $N_{k(x_i)}$ визначає набір k найближчих сусідів до x_i.

3. Support Vector Machine (svc):

$$\hat{y}_i^{(svc)} = \text{sign}(\sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i) + b)$$

де αj - коефіцієнти, які відповідають опорним векторам, b - зміщення у рішенні, а K - ядрова функція.

Мета-класифікатор (svc):

$$\hat{y}_i^{(meta)} = f_{meta}(\hat{Y}_i; \theta_{meta})$$

де $\hat{Y}_i = [\hat{y}_i^{(xgb)}, \hat{y}_i^{(knn)}, \hat{y}_i^{(svc)}]$ є вектором, який містить прогнози від кожного базового класифікатора для i-го елемента, а *fmeta* - функція прогнозування мета-класифікатора, що навчається на цих прогнозах.

Функція втрат:

$$L(Y, \hat{Y}^{(meta)}) = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(meta)})$$

де l - відповідна функція втрати, що міряє розбіжність між фактичними мітками Y та прогнозами мета-класифікатора $\hat{Y}^{(meta)}$.

Ці рівняння висвітлюють, як кожен класифікатор вносить вклад у загальний прогноз стекінг-ансамблю, а мета-класифікатор інтегрує ці прогнози для покращення загальної точності моделі.

2.4 Алгоритм прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання

Алгоритм прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання можна науково описати як послідовність етапів, спрямованих на розробку високоточної моделі, що

інтегрує різноманітні передбачувальні моделі через техніку стекінгу. Ось детальний науковий опис алгоритму:

Крок 1. Імпорт необхідних бібліотек. На цьому етапі здійснюється завантаження всіх бібліотек та модулів Python, необхідних для обробки даних, статистичного аналізу та машинного навчання. Зокрема, використовуються бібліотеки як pandas, numpy, scikit-learn, і matplotlib.

Крок 2. Підготовка даних. Цей крок включає завантаження датасету про серцеві захворювання, їх первинну обробку для видалення пропущених значень, нормалізацію даних та розбиття на тренувальні та тестові набори. Важливим елементом є також вибір релевантних ознак, які мають високу прогностичну цінність для серцевих захворювань.

Крок 3. Реалізація ансамблювання за допомогою стекінгу. Опис методу стекінгу, який полягає у комбінуванні прогнозів декількох базових алгоритмів машинного навчання для формування більш точного кінцевого прогнозу. Стекінг використовує мета-алгоритм, який на основі прогнозів базових моделей навчається робити кінцевий прогноз.

Крок 4. Налаштування стекінг-ансамблю. Ініціалізація базових моделей, таких як логістична регресія, випадковий ліс, і градієнтний бустінг, та мета-класифікатора. Налаштування параметрів цих моделей для оптимізації процесу навчання та зменшення перенавчання.

Крок 5. Навчання стекінг-ансамблю. Процес тренування базових моделей на тренувальних даних, їх валідація та збір прогнозів. Навчання мета-класифікатора на основі отриманих прогнозів базових моделей.

Крок 6. Оцінка моделі. Використання навченого ансамблю для оцінки тестових даних. Аналіз отриманих результатів з використанням матриці помилок, точності, чутливості, специфічності та інших метрик оцінки моделі. Циклічне уточнення та повторне тренування моделі за потребою.

Крок 7. Кінець. Завершення процесу після досягнення задовільного рівня точності або вичерпання визначених ресурсів для навчання.

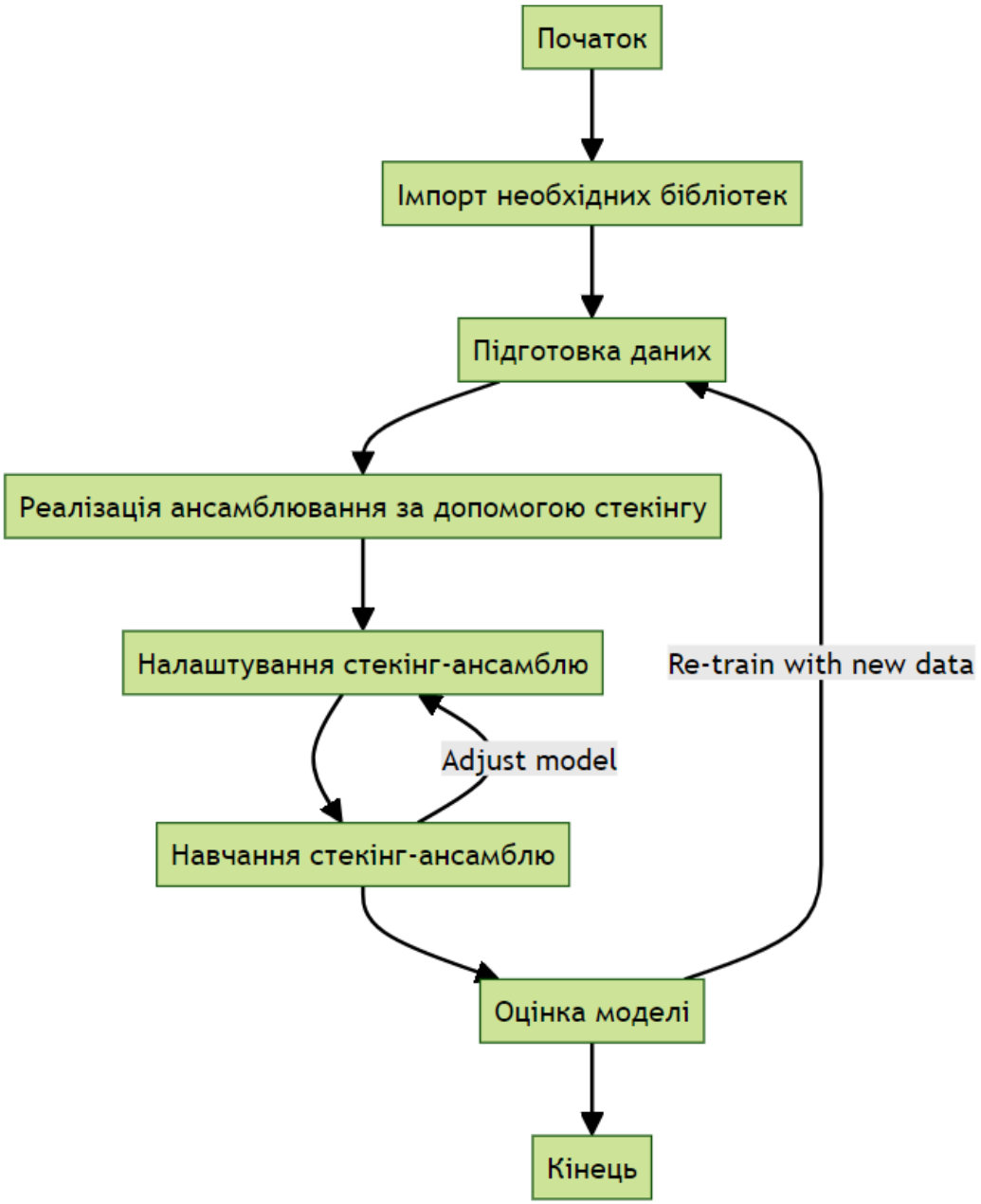


Рисунок 2.2 - Алгоритм прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання

Цей алгоритм демонструє використання складних машинних алгоритмів в сфері медичної діагностики та може бути адаптований або розширений для різних типів медичних даних та прогностичних завдань.

3 ПРОГРАМНО-ТЕХНОЛОГІЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Опис та аналіз даних

У даному дослідженні було проаналізовано набір даних, який містить інформацію про 303 пацієнтів з підозрою на серцеві захворювання. Змінні, що входять у набір даних, включають вік, стать, тип болю в грудях, артеріальний тиск у стані спокою, рівень холестерину, наявність цукру в крові натщесерце, результатах електрокардіографії в спокої, максимальній частоті серцебиття, наявності стенокардії при фізичному навантаженні, депресії ST, нахилі ST-сегмента, кількості основних судин, забарвленні талію та наявності захворювання серця. Завдяки використанню методів візуалізації даних (див. Рисунок 3.1), було виявлено ключові тенденції та аномалії у розподілі цих змінних.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Рисунок 3.1 – Структура даних

Під час статистичного аналізу (див. Таблицю 3.1) було визначено основні описові статистики для кожної змінної. Середній вік учасників

становив 54.4 року, при цьому стандартне відхилення було 9.08 років. Чоловіків серед учасників було більше (68.3%). Найчастіше зустрічалися скарги на тип болю в грудях класифікації 1. Аналіз варіабельності та розподілу інших медичних показників дозволив виявити паттерни, що можуть вказувати на підвищений ризик серцевих захворювань. Ці результати дозволяють глибше зрозуміти фактори ризику та можуть бути використані для покращення стратегій профілактики та лікування серцевих захворювань у населення.

Таблиця 3.1 – Описова статистика

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303	303	303	303	303	303	303	303	303	303	303	303	303	303
mean	54,37	0,68	0,97	131,62	246,26	0,15	0,53	149,65	0,33	1,04	1,40	0,73	2,31	0,54
std	9,08	0,47	1,03	17,54	51,83	0,36	0,53	22,91	0,47	1,16	0,62	1,02	0,61	0,50
min	29	0	0	94	126	0	0	71	0	0	0	0	0	0
25%	47,5	0	0	120	211	0	0	133,5	0	0	1	0	2	0
50%	55	1	1	130	240	0	1	153	0	0,8	1	0	2	1
75%	61	1	2	140	274,5	0	1	166	1	1,6	2	1	3	1
max	77	1	3	200	564	1	2	202	1	6,2	2	4	3	1

Матриця кореляцій (Рисунок 3.2) між різними клінічними змінними, які були використані в аналізі серцевих захворювань у 303 пацієнтів. Кожна клітина матриці відображає кореляцію між парами змінних, де кольори варіюються від синього (негативна кореляція) до червоного (позитивна кореляція). Наприклад, сильна червона зона між змінними "cp" (тип болю в грудях) та "target" (наявність серцевого захворювання) свідчить про високий рівень позитивної кореляції, що може означати, що тип болю в грудях є значимим індикатором ризику серцевих захворювань. Такі кореляції дозволяють визначити ключові змінні для подальшого глибокого аналізу та розробки медичних втручань.

З іншого боку, область з негативною кореляцією, така як між "thalach" (максимальна частота серцебиття) та "age" (вік), демонструє зниження частоти серцебиття зі збільшенням віку, що є очікуваною тенденцією. Ці знахідки,

відображені на рисунку, надають цінну інформацію для розуміння зв'язків між фізіологічними параметрами та їх впливом на ризик серцевих захворювань. Детальніше проаналізувавши ці кореляції, можна краще зрозуміти механізми розвитку захворювань та оптимізувати підходи до їх діагностики та лікування.

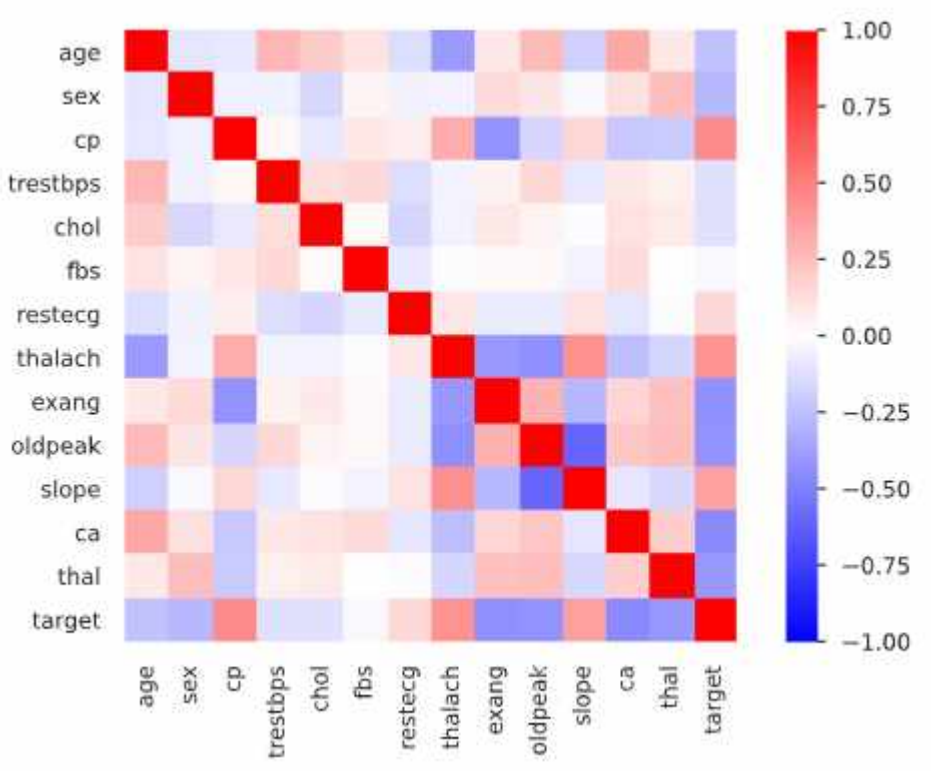


Рисунок 3.2 – Кореляційна матриця

3.2 Підготовка моделі та порівняльний аналіз

У процесі підготовки моделі для дослідження наявності серцевих захворювань дані були підготовлені за стандартною процедурою машинного навчання, що включала видалення цільової змінної з основного набору даних, нормалізацію даних та розділення на навчальну та тестову вибірки. Важливою кроком було забезпечення рівного розподілу даних між навчальним та тестовим наборами для запобігання проблеми дисбалансу даних. Розподіл навчальних даних показав, що кількість пацієнтів з серцевими

захворюваннями (131) та без них (111) була достатньо збалансована, що важливо для об'єктивної оцінки моделей.

Після підготовки і нормалізації даних було використано ряд алгоритмів машинного навчання для прогнозування серцевих захворювань. Алгоритми, такі як логістична регресія, наївний Баєсів класифікатор, класифікатор випадкового лісу, та інші були застосовані для визначення їх ефективності на тестових даних. Результати кожного класифікатора були оцінені за допомогою матриці плутанини та показників точності, що дозволило визначити найбільш ефективні методи в контексті даного набору даних.

Окрім порівняння базових показників точності, було проведено аналіз важливості змінних за допомогою технік машинного навчання. Такий аналіз допомагає виявити найважливіші ознаки, які впливають на прогнозування серцевих захворювань, зокрема, було використано модель екстремального градієнтного бустингу для оцінки важливості кожної змінної. Результати цього аналізу були представлені у вигляді графіка, що візуально демонструє рівень впливу кожної ознаки на результати моделі (Рисунок 3.3).

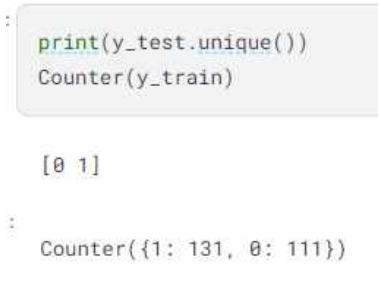


Рисунок 3.3 - Результати моделі

У даному розділі дослідження розглядаються різні алгоритми машинного навчання з метою визначення найточнішого прогнозування наявності серцевих захворювань. Використовуються такі моделі як логістична регресія, наївний Баєс, класифікатор випадкового лісу, екстремальний градієнтний бустинг, метод найближчих сусідів, дерево рішень та підтримуючі векторні машини. Кожна з цих моделей застосовується до одного набору даних, щоб визначити, яка з них найефективніше виявляє потенційні

серцеві захворювання, використовуючи статистичні метрики для оцінки точності кожного алгоритму.

Логістична регресія, застосована до набору даних для прогнозування серцевих захворювань, показала загальну точність 85.25%. Матриця плутанини (Рисунок 3.4) вказує, що модель коректно ідентифікувала 31 випадок наявності захворювання при 3 помилках (false negatives) та коректно визначила 21 випадок відсутності захворювання при 6 помилках (false positives). Згідно зі звітом класифікації, точність для випадків без захворювання складає 0.88, а для випадків з захворюванням — 0.84, що демонструє високу здатність моделі розрізняти стани здоров'я.

```
confussion matrix
[[21  6]
 [ 3 31]]


Accuracy of Logistic Regression: 85.24590163934425

              precision    recall  f1-score   support

           0       0.88      0.78      0.82        27
           1       0.84      0.91      0.87        34

    accuracy                           0.85        61
   macro avg       0.86      0.84      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Рисунок 3.4 – Результати Logistic Regression

Модель наївного Баєса, застосована для прогнозування серцевих захворювань, продемонструвала загальну точність 85.25%. Аналіз матриці плутанини показує, що модель ефективно виявила 31 випадок наявності захворювання з 3 помилково негативними результатами та визначила 21 випадок відсутності захворювання з 6 помилково позитивними результатами. Згідно зі звітом про класифікацію, точність діагностики відсутності захворювання складає 0.88, а точність діагностики наявності захворювання — 0.84, що відображає досить високу чутливість і специфічність моделі у виявленні серцевих захворювань (Рисунок 3.5).

Модель класифікатора випадкового лісу продемонструвала загальну точність у розмірі 85.25% у прогнозуванні наявності серцевих захворювань. Матриця плутанини відображає, що модель успішно ідентифікувала 30 випадків наявності захворювання з 4 помилково негативними результатами та визначила 22 випадки відсутності захворювання з 5 помилково позитивними результатами. Згідно зі звітом про класифікацію, модель має точність визначення відсутності захворювання 0.85 та точність визначення наявності захворювання 0.86, що свідчить про добре збалансовані показники чутливості та специфічності моделі у виявленні станів (Рисунок 3.6).

```
confussion matrix
[[21  6]
 [ 3 31]]


Accuracy of Naive Bayes model: 85.24590163934425

              precision    recall  f1-score   support

           0       0.88      0.78      0.82        27
           1       0.84      0.91      0.87        34

    accuracy                           0.85        61
   macro avg       0.86      0.84      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Рисунок 3.5 – Результати Naive Bayes

```
confussion matrix
[[22  5]
 [ 4 30]]


Accuracy of Random Forest: 85.24590163934425

              precision    recall  f1-score   support

           0       0.85      0.81      0.83        27
           1       0.86      0.88      0.87        34

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Рисунок 3.6 – Результати Random Forest

Модель Extreme Gradient Boost, налаштована з використанням глибоких налаштувань і кількісних оцінок, продемонструвала високу точність у 90.16% у прогнозуванні наявності серцевих захворювань. Матриця плутанини показує, що модель ефективно ідентифікувала 31 випадок наявності захворювання з лише 3 помилково негативними результатами, та точно визначила 24 випадки відсутності захворювання з 3 помилково позитивними результатами. Згідно з звітом про класифікацію, модель виявилася високо ефективною з балансом точності та відгуків між класами, забезпечуючи однаково високі показники для пацієнтів із захворюванням та без нього, що свідчить про її надійність та ефективність у диференціюванні станів здоров'я (Рисунок 3.7).

```
confussion matrix
[[24  3]
 [ 3 31]]


Accuracy of Extreme Gradient Boost: 90.1639344262295
```

	precision	recall	f1-score	support
0	0.89	0.89	0.89	27
1	0.91	0.91	0.91	34
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

Рисунок 3.7 – Результати Extreme Gradient Boost

Метод класифікації найближчих сусідів (K-NeighborsClassifier) був застосований для прогнозування наявності серцевих захворювань, досягнувши точності 88.52%. Аналіз матриці плутанини вказує на те, що модель ідентифікувала 30 випадків наявності захворювання при 4 помилково негативних результатів і визначила 24 випадки відсутності захворювання при 3 помилково позитивних результатів. Згідно зі звітом про класифікацію, модель має високі показники точності та відгуків, зокрема, точність діагностики відсутності захворювання складає 0.86, а точність діагностики

наявності захворювання — 0.91, що підтверджує її ефективність у виявленні різних станів здоров'я (Рисунок 3.8).

Дерево рішень (DecisionTreeClassifier), налаштоване з використанням критерію ентропії і максимальною глибиною 6 рівнів, показало точність у 81.97% у прогнозуванні наявності серцевих захворювань. Матриця плутанини виявила, що модель коректно ідентифікувала 27 випадків наявності захворювання з 7 помилково негативними результатами і точно визначила 23 випадки відсутності захворювання з 4 помилково позитивними результатами. Згідно зі звітом про класифікацію, модель продемонструвала вищу точність (0.87) для ідентифікації пацієнтів з захворюванням порівняно з точністю (0.77) для випадків без захворювань, підкреслюючи її ефективність, але із певною схильністю до помилково негативних результатів у виявленні станів здоров'я (Рисунок 3.9).

```
confussion matrix
[[24  3]
 [ 4 30]]


Accuracy of K-NeighborsClassifier: 88.52459016393442

              precision    recall  f1-score   support

           0       0.86      0.89      0.87        27
           1       0.91      0.88      0.90        34

    accuracy                           0.89        61
   macro avg       0.88      0.89      0.88        61
weighted avg       0.89      0.89      0.89        61
```

Рисунок 3.8 – Результати K-NeighborsClassifier

```
confussion matrix
[[23  4]
 [ 7 27]]


Accuracy of DecisionTreeClassifier: 81.9672131147541

              precision    recall  f1-score   support

           0       0.77      0.85      0.81        27
           1       0.87      0.79      0.83        34

    accuracy                           0.82        61
   macro avg       0.82      0.82      0.82        61
weighted avg       0.82      0.82      0.82        61
```

Рисунок 3.9 – Результати DecisionTreeClassifier

Класифікатор підтримуючих векторів (Support Vector Classifier) з ядром RBF та параметром регуляризації C=2 показав точність у 88.52% у прогнозуванні наявності серцевих захворювань. Матриця плутанини показує, що модель ефективно визначила 31 випадок наявності захворювання з лише 3 помилково негативними результатами і коректно ідентифікувала 23 випадки відсутності захворювання з 4 помилково позитивними результатами. Звіт про класифікацію свідчить про високу точність (0.89) і відгук (0.91) моделі у виявленні пацієнтів з захворюваннями, що підтверджує її здатність ефективно розрізняти між станами здоров'я (Рисунок 3.10).

```
confussion matrix
[[23  4]
 [ 3 31]]


Accuracy of Support Vector Classifier: 88.52459016393442

              precision    recall  f1-score   support

           0       0.88      0.85      0.87        27
           1       0.89      0.91      0.90        34

    accuracy                           0.89        61
   macro avg       0.89      0.88      0.88        61
weighted avg       0.89      0.89      0.88        61
```

Рисунок 3.10 – Результати Support Vector Classifier

На барплоті (Рисунок 3.11) відображено важливість особливостей за даними, отриманими з моделі екстремального градієнтного бустингу. Найважливішими факторами, що впливають на прогнозування наявності серцевих захворювань, виявилися 'thal' (тип таласемії), 'ca' (кількість великих судин, заблокованих кальцієм) і 'slope' (нахил піку серцевого імпульсу). Це підкреслює значення цих параметрів у діагностиці станів серцевої системи, що може бути корисним для подальших медичних досліджень і клінічної практики.

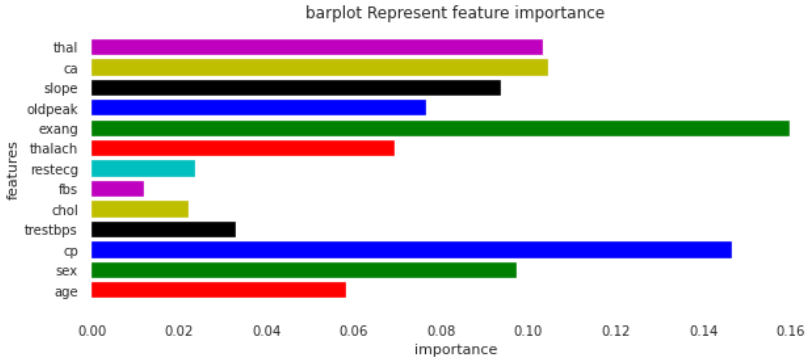


Рисунок 3.11 – Відображення важливості функції

Крива (Рисунок 3.12) робочих характеристик приймача (ROC) для різних моделей машинного навчання показує, як кожна модель здатна відрізняти між класами (наявність або відсутність захворювання) з різними порогами класифікації. Моделі, такі як екстремальний градієнтний бустинг та підтримуючий векторний класифікатор, продемонстрували високу точність у відношенні істинно позитивних та хибно позитивних ставок, що свідчить про їхню високу діагностичну ефективність порівняно з іншими моделями. Це забезпечує цінну інформацію для вибору найбільш відповідного методу аналізу в залежності від клінічних вимог та доступних даних.

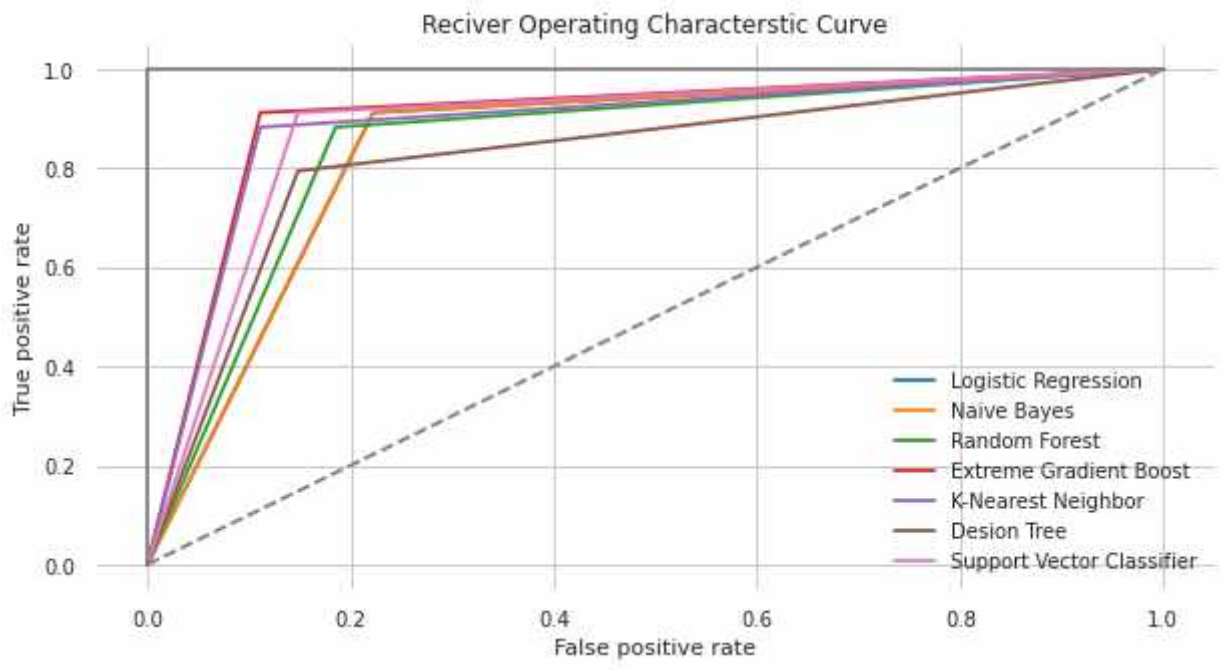


Рисунок 3.12 – Крива робочих характеристик приймача (ROC)

3.3 Оцінка ефективності моделей

Далі проведено порівняння ефективності семи різних моделей машинного навчання для прогнозування наявності серцевих захворювань. До аналізу були включені моделі: логістична регресія, наївний Баєс, випадковий ліс, екстремальний градієнтний бустинг, метод найближчих сусідів, дерево рішень та підтримуючі векторні машини. Точність кожної з моделей була оцінена на основі тестової вибірки, результати чого відображені на рисунку 3.13 (барплот).

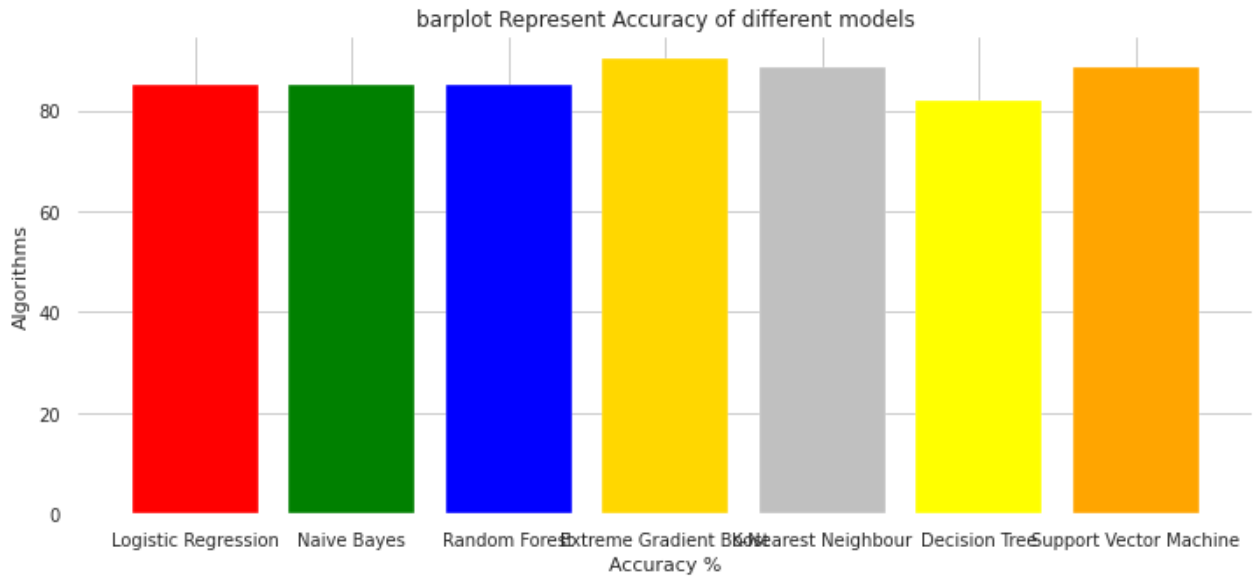


Рисунок 3.13 - Точність різних моделей

Згідно з результатами, найбільш ефективною моделлю виявилася модель екстремального градієнтного бустингу з точністю 90.16%, що свідчить про її високу здатність точно класифікувати випадки наявності та відсутності захворювань. На другому місці знаходяться моделі K-Nearest Neighbour та Support Vector Machine, які показали точність 88.52%. Ці моделі також демонструють високий рівень точності, що робить їх корисними для медичних діагностичних застосувань.

Моделі логістичної регресії, наївного Баєса та випадкового лісу показали однакову точність на рівні 85.25%. Ці результати вказують на стабільну здатність цих алгоритмів до класифікації, хоча їх ефективність дещо нижча порівняно з моделями, які посіли перші місця. Водночас, модель дерева рішень показала найнижчу точність у 81.97%, що може бути обумовлено її меншою здатністю до узагальнення даних.

Барплот (див. Рисунок 3.13) чітко демонструє різницю в точності між усіма моделями, підкреслюючи переваги і недоліки кожного підходу. Ефективність екстремального градієнтного бустингу та підтримуючих векторних машин особливо виділяється, вказуючи на їхню вищу здатність до точного прогнозування на основі даних про пацієнтів.

3.4 Ансамблювання для підвищення точності моделі

У рамках збільшення точності прогнозування наявності серцевих захворювань було застосовано техніку ансамблювання, зокрема стекінг. Стекінг, або стекова узагальнення, є методом ансамблю машинного навчання, який використовує мета-алгоритм навчання для визначення найкращого способу комбінування прогнозів від двох або більше базових алгоритмів машинного навчання. Базовий рівень зазвичай складається з різних алгоритмів навчання, тому ансамблі стекінгу часто є гетерогенними.

У цьому дослідженні використано StackingCVClassifier, який інтегрує моделі Extreme Gradient Boosting (XGB), K-Nearest Neighbors (KNN), і Support Vector Machine (SVC) як базові класифікатори, використовуючи SVC як мета-класифікатор. Модель (Рисунок 3.14) була навчена на тренувальній вибірці та перевірена на тестовій вибірці.

```
confussion matrix
[[24  3]
 [ 2 32]]


Accuracy of StackingCVClassifier: 91.80327868852459
```

	precision	recall	f1-score	support
0	0.92	0.89	0.91	27
1	0.91	0.94	0.93	34
accuracy			0.92	61
macro avg	0.92	0.92	0.92	61
weighted avg	0.92	0.92	0.92	61

Рисунок 3.14 – Результати StackingCVClassifier

Результати прогнозування моделі стекінгу показали високу точність у 91.80%, що є покращенням порівняно з окремими моделями, розглянутими раніше. Матриця плутанини для цієї моделі показала, що з 27 випадків відсутності захворювання, 24 були правильно ідентифіковані, тоді як 3 випадки були неправильно класифіковані. З 34 випадків наявності захворювання, 32 були точно визначені, а 2 — неправильно.

Звіт класифікації вказує на високі показники точності та відгуку для обох класів: точність ідентифікації відсутності захворювання становила 0.92, ідентифікації наявності захворювання — 0.91. Це свідчить про високий рівень збалансованості між чутливістю та специфічністю моделі, що забезпечує її надійність у клінічному застосуванні.

Таким чином, стекінг виявився ефективним підходом до покращення точності прогнозування серцевих захворювань, демонструючи значний потенціал для подальшого застосування у медичних діагностичних системах. Використання гетерогенних ансамблів дозволяє поєднувати переваги різних алгоритмів, забезпечуючи більшу адаптивність та точність у виявленні захворювань.

Алгоритмічне забезпечення прогнозування серцевих захворювань, застосовуючи ансамблеві методи машинного навчання, революціонізує підходи до моніторингу та інтервенцій у сфері здоров'я серця. Використання таких комплексних технік як стекінг дозволяє вченим і клініцистам інтегрувати різноманітні дані з електрокардіограм, магнітно-резонансних та комп'ютерних томографій для створення точних прогнозувальних моделей. Цей процес не тільки підвищує точність діагностики, але й дозволяє персоналізувати лікування, спираючись на деталізовані біомаркери і генетичні дані, що значно знижує ризики невдалої терапії та покращує загальні результати лікування.

Архітектура модуля прогнозування серцевих захворювань, що включає збір даних, їхню валідацію, попередню обробку, вибір ознак, тренування

моделей та врешті-решт впровадження моделей у клінічну практику, є взірцем інтеграції передових технологій у медичну індустрію. Результатом цієї інтеграції є покращення раннього виявлення захворювань, зменшення часу діагностики та оптимізація медичного обслуговування, забезпечуючи краще використання ресурсів і, що найважливіше, забезпечення вищої якості життя пацієнтів.

Програмно-технологічне забезпечення аналізу даних про серцеві захворювання виявляє значну ефективність в ідентифікації та прогнозуванні ризиків захворювань на серце. Детальний аналіз набору даних, що включає комплексні медичні показники, дозволяє виявити критичні змінні, які можуть впливати на здоров'я серцево-судинної системи. Використання статистичних методів і візуалізація даних забезпечують глибоке розуміння взаємозв'язків між різними параметрами та їхнім впливом на стан серцевої функції. Такий підхід сприяє розробці цільових стратегій для ефективної профілактики та лікування, заснованих на науково обґрунтованих даних.

Результати аналізу кореляції між клінічними показниками відіграють ключову роль у формуванні підходів до прогнозування серцевих захворювань. Відкриття взаємозв'язків, таких як позитивна кореляція між типом болю в грудях та наявністю серцевих захворювань, надають потенційні напрями для подальших досліджень та розробки діагностичних інструментів. Оцінка впливу цих змінних допомагає вченим і лікарям уточнити методи оцінки ризику та вибрати оптимальне лікування для пацієнтів, що в свою чергу може знизити загальну смертність від серцевих недуг.

ВИСНОВКИ

Сучасні підходи в діагностиці та лікуванні серцевих захворювань, зокрема інтеграція машинного навчання та штучного інтелекту, зазнали значних інновацій, що відкривають нові перспективи для поліпшення медичного догляду. Використання передових діагностичних технологій і методів, таких як ЕКГ, МРТ, комп'ютерна томографія, разом з розвитком алгоритмів машинного навчання дозволяє точніше ідентифікувати ризики і забезпечує раннє виявлення хвороб, значно покращуючи профілактику і терапевтичні стратегії. Автоматизація рутинних процесів і персоналізація медицини сприяють не тільки оптимізації лікування, але й значному зниженню смертності від серцево-судинних захворювань. Такий комплексний та інноваційний підхід підкреслює важливість продовження наукових досліджень і розробки нових технологій, що мають потенціал радикально трансформувати кардіологію та загальну практику медичного обслуговування.

Алгоритмічне забезпечення прогнозування серцевих захворювань з використанням ансамблевих методів машинного навчання демонструє значний потенціал у підвищенні точності діагностики та профілактики серцевих захворювань. Впровадження архітектури прогнозування, що охоплює збір, перевірку та попередню обробку даних, вибір ознак, тренування моделей та їх розгортання, забезпечує комплексний підхід до аналізу медичних даних. Стекінг-ансамблі, що поєднують прогнози від кількох базових моделей, дозволяють мета-класифікатору об'єднувати ці прогнози для отримання більш точних результатів. Оптимізація параметрів моделей та використання методів перехресної перевірки зменшують ризик перенавчання та підвищують надійність системи. Завдяки таким технологіям, прогнозування серцевих захворювань може стати ефективним інструментом в медичній практиці, значно зменшуючи ризики для здоров'я пацієнтів та підвищуючи якість медичного обслуговування.

Програмно-технологічне забезпечення аналізу даних про серцеві захворювання виявилося надзвичайно ефективним в ідентифікації та прогнозуванні ризиків серцевих недуг. У процесі дослідження було проаналізовано детальний набір даних, що включає широкий спектр клінічних показників, таких як вік, стать, тип болю в грудях, артеріальний тиск, рівень холестерину та інші. Використання статистичних методів та візуалізація даних дозволили виявити критичні змінні та ключові тенденції, що впливають на стан серцево-судинної системи. Кореляційний аналіз між клінічними змінними виявив важливі взаємозв'язки, такі як позитивна кореляція між типом болю в грудях та наявністю серцевих захворювань, що може вказувати на значимість цих показників у прогнозуванні. Завдяки таким підходам стало можливим розробити цільові стратегії для ефективної профілактики та лікування серцевих захворювань, що базуються на науково обґрунтованих даних. Це дослідження підтверджує важливість інтеграції передових технологій аналізу даних у медичну практику, що сприяє покращенню діагностики, підвищенню якості медичного обслуговування та зниженню смертності від серцевих недуг.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Prasad, D. D., Pratyusha, Y., Jaswitha, A. J. H. S., & ін. (2024). FBGN: A Fusion Based Genetic PSO Network for Heart Attack Prediction. *Journal of Management and Engineering Integration*. Retrieved from http://dmamerabek.org/index.php/dmame/article/view/83
2. Bhattacharya, N., & Kumar, P. (2024). Integrating Healthcare Analytics to Improve Diabetes Management and Prevent Heart Attacks: A Data-Driven Approach. *Research Square*. Retrieved from https://www.researchsquare.com/article/rs-4310669/latest.pdf
3. Islam, M. A., Majumder, M. Z. H., Miah, M. S., & ін. (2024). Precision Healthcare: A Deep Dive into Machine Learning Algorithms and Feature Selection Strategies for Accurate Heart Disease Prediction. *Computers in Biology and Medicine*. Retrieved from https://www.sciencedirect.com/science/article/pii/S001048252400516X
4. Topol, E. J. (2024). AI-enabled opportunistic medical scan interpretation. *The Lancet*. Retrieved from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00924-3/abstract
5. Booba, B. (2024). Predict Diabetes Healthcare Analytics Using Hybrid Gradient Boosting Machine Learning Model. *Educational Administration: Theory and Practice*. Retrieved from http://www.kuey.net/index.php/kuey/article/view/3371
6. Bhattacharya, N., & Kumar, P. (2024). Integrating Healthcare Analytics to Improve Diabetes Management and Prevent Heart Attacks: A Data-Driven Approach. Research Square. Retrieved from https://www.researchsquare.com/article/rs-4310669/latest.pdf
7. Islam, M. A., Majumder, M. Z. H., Miah, M. S., & ін. (2024). Precision Healthcare: A Deep Dive into Machine Learning Algorithms and Feature

Selection Strategies for Accurate Heart Disease Prediction. Computers in Biology and Medicine. Retrieved from https://www.sciencedirect.com/science/article/pii/S001048252400516X

8. Wang, M., Li, G., Yang, Y., Feng, Y., & Li, Y. (2024). Automated analysis of fetal heart rate baseline/acceleration/deceleration using MTU-Net3+ model. Biomedical Engineering. Retrieved from https://link.springer.com/article/10.1007/s13534-024-00388-x
9. Kaushal, P., & Singh, S. (2024). Encoder-Based Universal Transformer for Survival Analysis—A Case Study on Right Censored Heart Failure Data. Arabian Journal for Science and Engineering. Retrieved from https://link.springer.com/article/10.1007/s13369-024-09093-4
10. Miller, R. J. H., Shanbhag, A., Michalowska, A. M., & ін. (2024). Deep Learning–Enabled Quantification of 99mTc-Pyrophosphate SPECT/CT for Cardiac Amyloidosis. Journal of Nuclear Medicine. Retrieved from https://jnm.snmjournals.org/content/early/2024/05/09/jnumed.124.267542.abstract
11. Комар М.П., Саченко А.О., Васильків Н.М., Гладій Г.М., Коваль В.С., Ліп’яніна-Гончаренко Х.В. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за першим (бакалаврським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 52 с.
12. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.

ДОДАТОК А

ПСЕВДОКОД АЛГОРИТМУ

```
Algorithm HeartDiseasePrediction
  Import libraries

  LoadData:
    Read heart.csv into data

  DisplayBasicInfo:
    Display first few rows of data
    Display info of data
    Display missing values in data
    Describe data

  GenerateProfileReport:
    Generate and display profile report using pandas_profiling

  PrepareDataForModel:
    Extract target variable 'target' into y
    Drop 'target' from data and assign rest to X
    Split data into X_train, X_test, y_train, y_test with 20% test size and random state 0
    Initialize and apply StandardScaler to X_train and X_test

  ModelTrainingAndEvaluation:
    Define and train models (Logistic Regression, Naive Bayes, Random Forest, XGBoost, KNN, Decision Tree,
SVM)
    Predict and evaluate each model using confusion matrix, accuracy score, and classification report
    Plot feature importance for XGBoost

  ModelComparison:
    Create DataFrame with model names and their accuracies
    Plot bar graph comparing model accuracies

  ImplementEnsembling:
    Define StackingCVClassifier with XGBoost, KNN, SVM as base classifiers and SVM as meta-classifier
    Train and predict using StackingCVClassifier
    Evaluate StackingCVClassifier using confusion matrix and classification report

  PlotROC:
    Calculate ROC curves for all models
    Plot ROC curves comparing all models

End Algorithm
```

ДОДАТОК Б

КОД ДЛЯ РЕАЛІЗАЦІЇ

```
import pandas as pd
import numpy as np
#visualisation
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
#EDA
from collections import Counter
import pandas_profiling as pp
# data preprocessing
from sklearn.preprocessing import StandardScaler
# data splitting
from sklearn.model_selection import train_test_split
# data modeling
from sklearn.metrics import confusion_matrix,accuracy_score,roc_curve,classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
#ensembling
from mlxtend.classifier import StackingCVClassifier

# %% [code]
data = pd.read_csv('../input/heart-disease-uci/heart.csv')
data.head()

# %% [code]
data.info()

# %% [markdown]
# ### **Missing Value Detection**

# %% [code]
data.isnull().sum()

# %% [markdown]
# ### **Descriptive statistics**

# %% [code]
data.describe()

# %% [markdown]
# ## **EDA**

# %% [code]
pp.ProfileReport(data)

# %% [markdown]
# ## **Model prepration**
```

```
# %% [code]
y = data["target"]
X = data.drop('target',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state = 0)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# %% [markdown]
# **Before applying algorithm we should check whether the data is equally splitted or not, because if data is not splitted equally it will cause for data imbalacing problem**

# %% [code]
print(y_test.unique())
Counter(y_train)

# %% [markdown]
# ## **ML models**
#
# Here I take different machine learning algorithm and try to find algorithm which predict accurately.
#
# 1. Logistic Regression
# 2. Naive Bayes
# 3. Random Forest Classifier
# 4. Extreme Gradient Boost
# 5. K-Nearest Neighbour
# 6. Decision Tree
# 7. Support Vector Machine
#

# %% [code]
m1 = 'Logistic Regression'
lr = LogisticRegression()
model = lr.fit(X_train, y_train)
lr_predict = lr.predict(X_test)
lr_conf_matrix = confusion_matrix(y_test, lr_predict)
lr_acc_score = accuracy_score(y_test, lr_predict)
print("confussion matrix")
print(lr_conf_matrix)
print("\n")
print("Accuracy of Logistic Regression:",lr_acc_score*100,'\n')
print(classification_report(y_test,lr_predict))

# %% [code]

m2 = 'Naive Bayes'
nb = GaussianNB()
nb.fit(X_train,y_train)
nbpred = nb.predict(X_test)
nb_conf_matrix = confusion_matrix(y_test, nbpred)
nb_acc_score = accuracy_score(y_test, nbpred)
print("confussion matrix")
print(nb_conf_matrix)
print("\n")
print("Accuracy of Naive Bayes model:",nb_acc_score*100,'\n')
print(classification_report(y_test,nbpred))

# %% [code]
```

```
m3 = 'Random Forest Classfier'
rf = RandomForestClassifier(n_estimators=20, random_state=2,max_depth=5)
rf.fit(X_train,y_train)
rf_predicted = rf.predict(X_test)
rf_conf_matrix = confusion_matrix(y_test, rf_predicted)
rf_acc_score = accuracy_score(y_test, rf_predicted)
print("confussion matrix")
print(rf_conf_matrix)
print("\n")
print("Accuracy of Random Forest:",rf_acc_score*100,'\n')
print(classification_report(y_test,rf_predicted))

# %% [code]
m4 = 'Extreme Gradient Boost'
xgb = XGBClassifier(learning_rate=0.01, n_estimators=25, max_depth=15,gamma=0.6,
subsample=0.52,colsample_bytree=0.6,seed=27,
                reg_lambda=2, booster='dart', colsample_bylevel=0.6, colsample_bynode=0.5)
xgb.fit(X_train, y_train)
xgb_predicted = xgb.predict(X_test)
xgb_conf_matrix = confusion_matrix(y_test, xgb_predicted)
xgb_acc_score = accuracy_score(y_test, xgb_predicted)
print("confussion matrix")
print(xgb_conf_matrix)
print("\n")
print("Accuracy of Extreme Gradient Boost:",xgb_acc_score*100,'\n')
print(classification_report(y_test,xgb_predicted))

# %% [code]
m5 = 'K-NeighborsClassifier'
knn = KNeighborsClassifier(n_neighbors=10)
knn.fit(X_train, y_train)
knn_predicted = knn.predict(X_test)
knn_conf_matrix = confusion_matrix(y_test, knn_predicted)
knn_acc_score = accuracy_score(y_test, knn_predicted)
print("confussion matrix")
print(knn_conf_matrix)
print("\n")
print("Accuracy of K-NeighborsClassifier:",knn_acc_score*100,'\n')
print(classification_report(y_test,knn_predicted))

# %% [code]
m6 = 'DecisionTreeClassifier'
dt = DecisionTreeClassifier(criterion = 'entropy',random_state=0,max_depth = 6)
dt.fit(X_train, y_train)
dt_predicted = dt.predict(X_test)
dt_conf_matrix = confusion_matrix(y_test, dt_predicted)
dt_acc_score = accuracy_score(y_test, dt_predicted)
print("confussion matrix")
print(dt_conf_matrix)
print("\n")
print("Accuracy of DecisionTreeClassifier:",dt_acc_score*100,'\n')
print(classification_report(y_test,dt_predicted))

# %% [code]
m7 = 'Support Vector Classifier'
svc =  SVC(kernel='rbf', C=2)
svc.fit(X_train, y_train)
svc_predicted = svc.predict(X_test)
```

```
svc_conf_matrix = confusion_matrix(y_test, svc_predicted)
svc_acc_score = accuracy_score(y_test, svc_predicted)
print("confussion matrix")
print(svc_conf_matrix)
print("\n")
print("Accuracy of Support Vector Classifier:",svc_acc_score*100,'\n')
print(classification_report(y_test,svc_predicted))

# %% [code]
imp_feature = pd.DataFrame({'Feature': ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
     'exang', 'oldpeak', 'slope', 'ca', 'thal'], 'Importance': xgb.feature_importances_})
plt.figure(figsize=(10,4))
plt.title("barplot Represent feature importance ")
plt.xlabel("importance ")
plt.ylabel("features")
plt.barh(imp_feature['Feature'],imp_feature['Importance'],color = 'rgbkymc')
plt.show()

# %% [code]
lr_false_positive_rate,lr_true_positive_rate,lr_threshold = roc_curve(y_test,lr_predict)
nb_false_positive_rate,nb_true_positive_rate,nb_threshold = roc_curve(y_test,nbpred)
rf_false_positive_rate,rf_true_positive_rate,rf_threshold = roc_curve(y_test,rf_predicted)
xgb_false_positive_rate,xgb_true_positive_rate,xgb_threshold = roc_curve(y_test,xgb_predicted)
knn_false_positive_rate,knn_true_positive_rate,knn_threshold = roc_curve(y_test,knn_predicted)
dt_false_positive_rate,dt_true_positive_rate,dt_threshold = roc_curve(y_test,dt_predicted)
svc_false_positive_rate,svc_true_positive_rate,svc_threshold = roc_curve(y_test,svc_predicted)


sns.set_style('whitegrid')
plt.figure(figsize=(10,5))
plt.title('Reciver Operating Characterstic Curve')
plt.plot(lr_false_positive_rate,lr_true_positive_rate,label='Logistic Regression')
plt.plot(nb_false_positive_rate,nb_true_positive_rate,label='Naive Bayes')
plt.plot(rf_false_positive_rate,rf_true_positive_rate,label='Random Forest')
plt.plot(xgb_false_positive_rate,xgb_true_positive_rate,label='Extreme Gradient Boost')
plt.plot(knn_false_positive_rate,knn_true_positive_rate,label='K-Nearest Neighbor')
plt.plot(dt_false_positive_rate,dt_true_positive_rate,label='Desion Tree')
plt.plot(svc_false_positive_rate,svc_true_positive_rate,label='Support Vector Classifier')
plt.plot([0,1],ls='--')
plt.plot([0,0],[1,0],c='.5')
plt.plot([1,1],c='.5')
plt.ylabel('True positive rate')
plt.xlabel('False positive rate')
plt.legend()
plt.show()

# %% [markdown]
# # **Model Evaluation**

# %% [code]
model_ev = pd.DataFrame({'Model': ['Logistic Regression','Naive Bayes','Random Forest','Extreme Gradient
Boost',
            'K-Nearest Neighbour','Decision Tree','Support Vector Machine'], 'Accuracy': [lr_acc_score*100,

nb_acc_score*100,rf_acc_score*100,xgb_acc_score*100,knn_acc_score*100,dt_acc_score*100,svc_acc_score
*100]})
model_ev
```

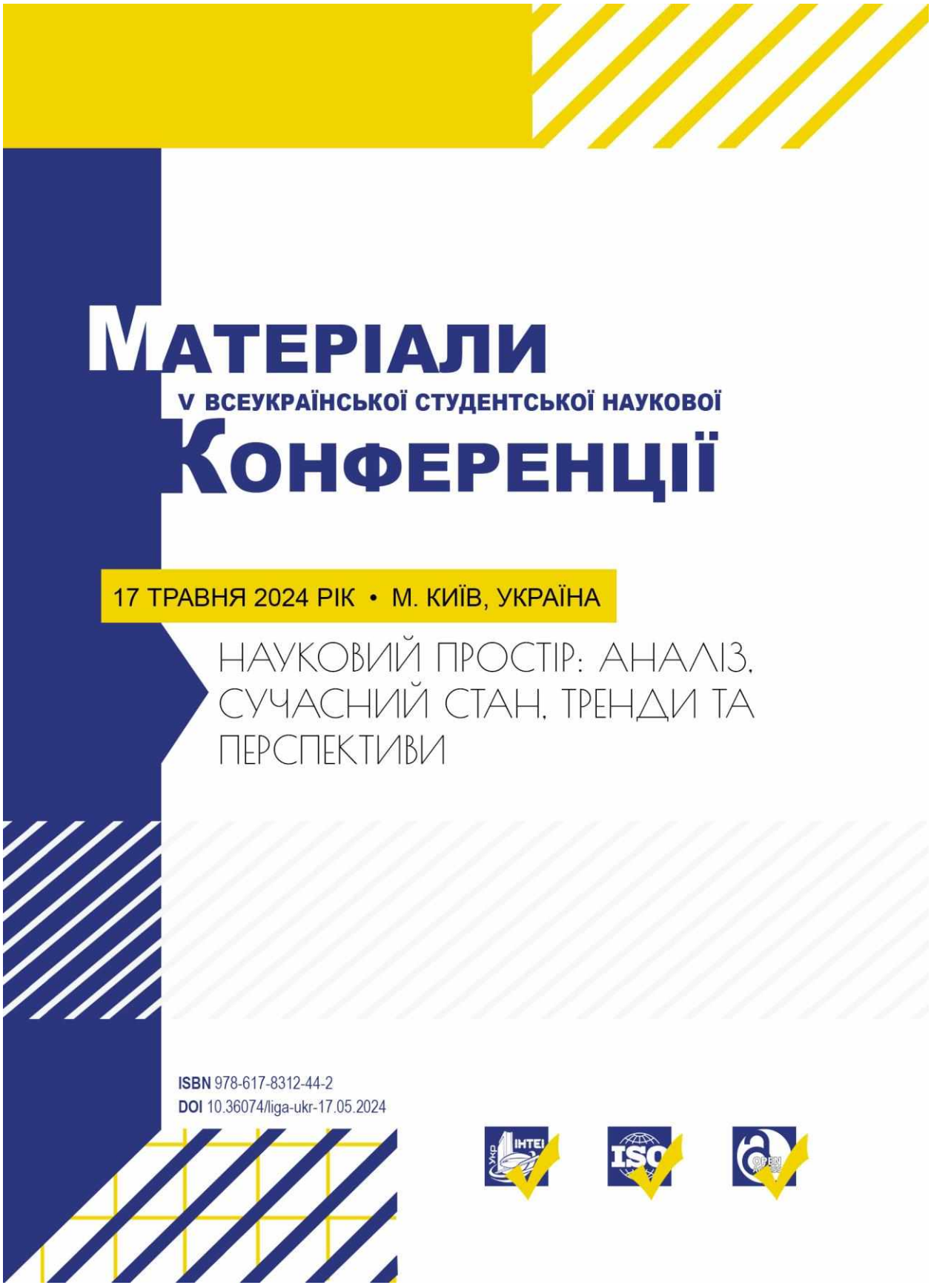
```
# %% [code]
colors = ['red','green','blue','gold','silver','yellow','orange',]
plt.figure(figsize=(12,5))
plt.title("barplot Represent Accuracy of different models")
plt.xlabel("Accuracy %")
plt.ylabel("Algorithms")
plt.bar(model_ev['Model'],model_ev['Accuracy'],color = colors)
plt.show()

# %% [markdown]
# ## **Ensembling**
#
# > In order to increase the accuracy of the model we use ensembling. Here we use stacking technique.
#
# ### About Stacking
#
# > Stacking or Stacked Generalization is an ensemble machine learning algorithm. It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms. The base level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous.The stacking ensemble is illustrated in the figure below
#
# > <img style="float: centre;" src="https://mlfromscratch.com/content/images/2020/01/image-2.png" width="400px"/>

# %% [code]
scv=StackingCVClassifier(classifiers=[xgb,knn,svc],meta_classifier= svc,random_state=42)
scv.fit(X_train,y_train)
scv_predicted = scv.predict(X_test)
scv_conf_matrix = confusion_matrix(y_test, scv_predicted)
scv_acc_score = accuracy_score(y_test, scv_predicted)
print("confussion matrix")
print(scv_conf_matrix)
print("\n")
print("Accuracy of StackingCVClassifier:",scv_acc_score*100,'\n')
print(classification_report(y_test,scv_predicted))
```

ДОДАТОК В

АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ



УДК 082:001
Н 34

Голова оргкомітету: Коренюк І.О.

Верстка: Зрада С.І.
Дизайн: Бондаренко І.В.

Рекомендовано до видання Вченою Радою Інституту науково-технічної інтеграції та співпраці. Протокол № 36 від 16.05.2024 року.

Конференцію зареєстровано Державною науковою установою «УкрІНТЕІ» в базі даних науково-технічних заходів України та інформаційному бюлетені «План проведення наукових, науково-технічних заходів в Україні» (Посвідчення №29 від 05.01.2024).

Н 34 **Науковий простір: аналіз, сучасний стан, тренди та перспективи**: матеріали V Всеукраїнської студентської наукової конференції, м. Київ, 17 травня, 2024 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2024. — 586 с.

ISBN 978-617-8312-44-2
DOI 10.62732/liga-ukr-17.05.2024

Викладено матеріали учасників V Всеукраїнської мультидисциплінарної студентської наукової конференції «Науковий простір: аналіз, сучасний стан, тренди та перспективи», яка відбулася 17 травня 2024 року у місті Київ, Україна.

УДК 082:001

ISBN 978-617-8312-44-2

СЕКЦІЯ 17.
ФІЗИКО-МАТЕМАТИЧНІ НАУКИ

Закалюк Павло, здобувач вищої освіти факультету
комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Констанкевич Ю.В., канд. техн. наук, доцент,
доцент кафедри інформаційно-обчислювальних систем і управління
Західноукраїнський національний університет, Україна

АРХІТЕКТУРА МОДУЛЯ ПРОГНОЗУВАННЯ СЕРЦЕВИХ ЗАХВОРЮВАНЬ

Прогнозування серцевих захворювань в сучасному світі викликає значний інтерес серед науковців та медичних фахівців, оскільки це дозволяє значно підвищити ефективність лікування та профілактики цих захворювань. Завдяки використанню передових технологій машинного навчання та збору великих обсягів медичних даних, можливо створити системи, здатні з високою точністю прогнозувати ризик розвитку серцевих захворювань у пацієнтів. Цей процес включає декілька критичних етапів, починаючи зі збору і перевірки даних, попередньої обробки, вибору відповідних ознак, тренування моделей і закінчуючи розгортанням готової моделі для використання в реальних умовах.

Прогнозування серцевих захворювань складається з декількох основних компонентів (Рисунок 1), які взаємодіють для забезпечення точного аналізу та прогнозування. Перший етап в системі - це збір даних, де здійснюється збір релевантної інформації, яка може включати клінічні дані, результати медичних обстежень, а також історію хвороб пацієнтів. Цей компонент також забезпечує перевірку зібраних даних на предмет їх повноти та достовірності, що є критично важливим для подальшого аналізу.

Наступний крок, попередня обробка даних, включає видалення неповних або аномальних записів, нормалізацію числових значень та розділення набору даних на тренувальні та тестові підмножини. Цей етап підготовки є фундаментальним для забезпечення того, що наступні моделі навчання машин будуть працювати з чистими та структурованими даними, що значно підвищує якість та надійність прогнозів.

Компонент вибір ознак відповідає за аналіз та вибір найбільш інформативних атрибутів з даних, які будуть використовуватися для тренування моделей. Цей процес включає статистичний аналіз та застосування алгоритмів вибору ознак, щоб визначити, які ознаки мають найсильніший вплив на результати прогнозування. Вибір правильних ознак збільшує точність моделі та ефективність навчання.

Навчання моделі є центральним компонентом системи, де використовуються обрані ознаки для тренування алгоритмів машинного навчання. В рамках цього компонента відбувається оптимізація параметрів моделі, включаючи налаштування гіперпараметрів та вибір найефективнішої моделі на основі перехресної перевірки та інших методів валідації.

Останній компонент, служба прогнозування, включає в себе розгортання тренованої моделі у виробниче середовище, де вона може надавати прогнози в реальному часі. Також цей компонент забезпечує моніторинг і оцінку продуктивності моделі, що дозволяє своєчасно виявляти та коригувати будь-які відхилення у точності

прогнозування.

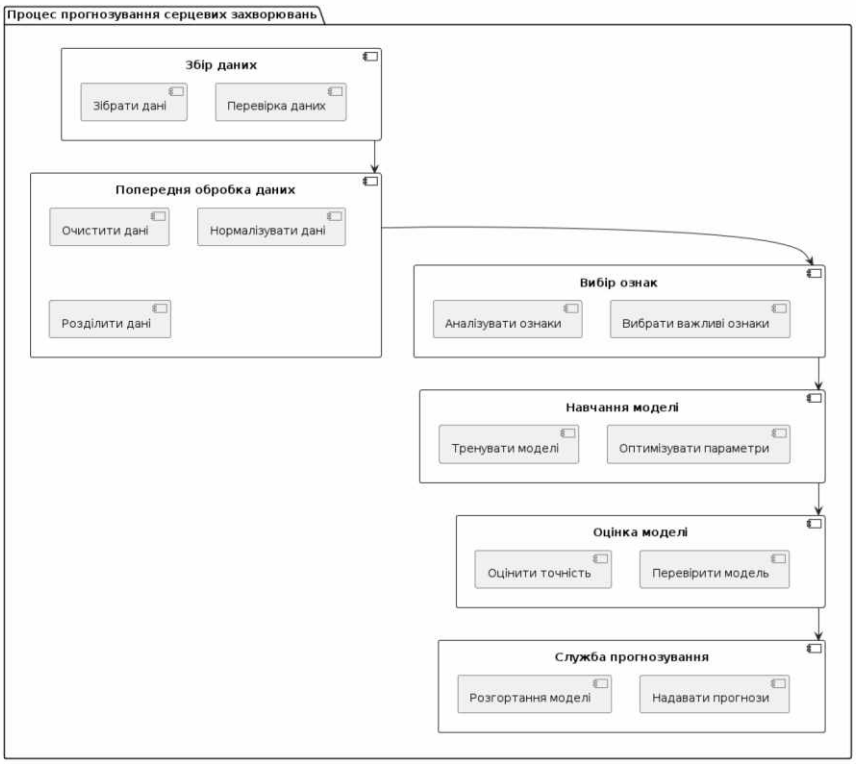


Рис. 1. **Архітектура [1] модуля прогнозування серцевих захворювань**

Архітектура системи прогнозування серцевих захворювань, представлена на рисунку 1, демонструє комплексний підхід до обробки та аналізу медичних даних для прогнозування. Кожен компонент системи відіграє ключову роль у забезпеченні точності та надійності прогнозу, де кінцевий компонент — служба прогнозування — забезпечує впровадження моделі в клінічну практику. Успішне використання такої системи може значно зменшити ризики для здоров'я пацієнтів, оптимізувати процеси медичного обслуговування та підвищити загальну якість життя популяції, що робить вивчення та розробку цих технологій надзвичайно важливою задачею сучасної медицини.

Список використаних джерел:

1. Архітектура програмного забезпечення для математичного моделювання на основі аналізу інтервальних даних з використанням хмарних технологій / М. Дивак та ін. Measuring and computing devices in technological processes. 2024. № 1. С. 125–139. URL: https://doi.org/10.31891/2219-9365-2024-77-15 (дата звернення: 13.05.2024).