

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ОДЕСЬКА ЮРИДИЧНА АКАДЕМІЯ»
ФАКУЛЬТЕТ КІБЕРБЕЗПЕКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інформаційних технологій

КВАЛІФІКАЦІЙНА РОБОТА

на тему:

**«МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПОКРАЩЕННЯ
ПРОГНОЗУВАННЯ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ»**

Виконав: студент 2 курсу

рівень: магістр

галузь знань 12 «Інформаційні
технології»

спеціальність 124 «Системний аналіз»

Ключник Роман Євгенович

Керівник: к.пед.н., доцент

Лобода Юлія Геннадіївна

Рецензент: к.пед.н., доцент

Логінова Наталія Іванівна

Одеса – 2024

АНОТАЦІЯ

Кваліфікаційна робота на тему «Методи інтелектуального аналізу даних для покращення прогнозування серцево-судинних захворювань» на здобуття другого (магістерського) рівня вищої освіти за спеціальністю 124 Системний аналіз, освітньою програмою: Аналіз та безпека даних, містить 26 рисунків, 55 літературних джерел за переліком посилань. Робота виконана на 67 сторінках загального тексту і 62 сторінках основного тексту.

Метою роботи є дослідження та аналіз факторів покращення прогнозування серцево-судинних захворювань з використанням методів інтелектуального аналізу даних.

Об'єктом дослідження є процес інтелектуального аналізу даних серцево-судинних захворювань.

Предметом дослідження виступають методи інтелектуального аналізу даних для покращення прогнозування серцево-судинних захворювань.

У кваліфікаційній роботі розглядаються підходи щодо використання методів інтелектуального аналізу даних для прогнозування серцево-судинних захворювань. Відмічено, що ефективність лікування серцево-судинних захворювань значною мірою залежить від точності прогнозу, своєчасної діагностики, комплексної профілактики та виявлення факторів ризику. Проаналізовано використання групи методів аналізу виживання для оцінки тривалості життя. Представлена реалізація моделі прогнозування серцево-судинних захворювань. Розрахована функція виживання, яка визначає можливість виживання протягом певного періоду часу. Проведено оцінку Каплана-Майєра для кривої виживання за виявленими ознаками. Встановлено фактори ризику серцево-судинних захворювань. Створено регресійну модель пропорційних ризиків Коксу для оцінки взаємозв'язку між декількома факторами ризику та часом виживання пацієнта.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ, ПРОГНОЗУВАННЯ, АНАЛІЗ ВИЖИВАННЯ, СЕРЦЕВО-СУДИННІ ЗАХВОРЮВАННЯ

ABSTRACT

Qualification work on the topic "Methods of intelligent data analysis to improve the prediction of cardiovascular diseases" for the second (master's) level of higher education in the specialty 124 System analysis, educational program: Data analysis and security, contains 26 figures, 55 literary sources according to the list of references. The work was completed on 67 pages of the general text and 62 pages of the main text.

The purpose of the work is research and analysis of factors for improving the prognosis of cardiovascular diseases using methods of intellectual data analysis.

The object of the study is the process of intellectual analysis of cardiovascular disease data.

The subject of research is the methods of intellectual data analysis to improve the prediction of cardiovascular diseases

The qualification work considers approaches to the use of methods of intellectual data analysis for predicting cardiovascular diseases. It was noted that the effectiveness of the treatment of cardiovascular diseases largely depends on the accuracy of the prognosis, timely diagnosis, comprehensive prevention and identification of risk factors. The use of a group of survival analysis methods to estimate life expectancy is analyzed. The implementation of the model for predicting cardiovascular diseases is presented. A survival function is calculated, which determines the possibility of survival during a certain period of time. A Kaplan-Meier estimate was performed for the survival curve based on the identified features. Risk factors for cardiovascular diseases have been established. A Cox proportional hazards regression model was created to assess the relationship between several risk factors and patient survival time.

INTELLECTUAL DATA ANALYSIS, METHODS OF INTELLECTUAL ANALYSIS,
FORECASTING, SURVIVAL ANALYSIS, CARDIOVASCULAR DISEASES

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ.....	6
ВСТУП.....	7
1. АНАЛІЗ СУЧАСНОГО СТАНУ ВИКОРИСТАННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ СЕРЦЕВО- СУДИННИХ ЗАХВОРЮВАНЬ	10
1.1. Аналіз предметної області.....	10
1.2. Методи інтелектуального аналізу даних	12
1.3. Процес інтелектуального аналізу медичних даних	17
Висновки до розділу 1.....	20
2. ГРУПА МЕТОДІВ SURVIVAL ANALYSIS ДЛЯ ПРОГНОЗУВАННЯ ТРИВАЛОСТІ ЖИТТЯ	21
2.1. Методи аналізу виживання.....	21
2.2. Криві виживання Каплана-Майєра та логарифмічний ранговий тест	27
2.3. Моделювання даних аналізу виживання за допомогою регресії Коксу	30
Висновки до розділу 2.....	34
3. ПРОГРАМНА РЕАЛІЗАЦІЯ ПРОГНОЗУВАННЯ СЕРЦЕВО СУДИННИХ ЗАХВОРЮВАНЬ.....	35
3.1. Розвідувальний аналіз даних.....	35
3.2. Оцінка Каплана-Майєра для кривої виживання	44
3.3. Аналіз відношення ризиків у моделі регресії Коксу	52
Висновки до розділу 3.....	54
ВИСНОВКИ.....	55
ПЕРЕЛІК ПОСИЛАНЬ	57
Додаток А. Графіки оцінки густини даних.....	63
Додаток Б. Результаті оцінки виживання Каплана-Майєра.....	64
Додаток В. Програмний код.....	65

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ССЗ – серцево-судинні захворювання

ССР – серцево-судинний ризик

ВООЗ – Всесвітня Організація Охорони Здоров'я

ІАД – інтелектуальний аналіз даних

ЕМК – електронна медична картка

k-NN – метод k-найближчих сусідів

SVM – метод опорних векторів

k-Means – метод k-середніх

DT – метод дерево рішень

БМ – байєсівська мережа

HR – відношення ризиків

АТ – артеріальний тиск

ФВ – фракція викиду

КФК – креатинінфосфокіназа

ІХС – ішемічна хвороба серця

EDA (Exploratory data analysis) – розвідувальний аналіз даних

ВСТУП

Серцево-судинні захворювання є одними з основних причин смертності у всьому світі. Впровадження швидкого та точного способу прогнозування серцево-судинних захворювань може значно покращити діагностику, лікування та комплексну профілактику захворювань.

Методи інтелектуального аналізу даних останніми роками набули великої популярності завдяки своїм можливостям аналізувати великі обсяги даних, виявляти складні закономірності. Використання таких методів у медичній сфері може допомогти ідентифікувати важливі прогностичні фактори та знайти нові зв'язки, які досі залишалися прихованими.

Важливість цього дослідження особливо проявляється у контексті зростання обсягу медичних даних, які накопичувалися останні роки. Враховуючи, що дані містять цінну інформацію про симптоми, фактори ризику, засоби лікування тощо, інтелектуальний аналіз може розкрити нові можливості для розуміння та прогнозування серцево-судинних захворювань. Точніші прогнози можуть покращити виявлення пацієнтів з високим ризиком серцево-судинних захворювань та призначення необхідного лікування на ранніх стадіях захворювання. Також це дозволяє знизити витрати та оптимізувати сферу медичного обслуговування. Тому тема кваліфікаційної роботи має велику актуальність.

Метою роботи є дослідження та аналіз факторів покращення прогнозування серцево-судинних захворювань з використанням методів інтелектуального аналізу даних.

Об'єктом дослідження є процес інтелектуального аналізу даних серцево-судинних захворювань.

Предметом дослідження виступають методи інтелектуального аналізу даних для покращення прогнозування серцево-судинних захворювань.

Для досягнення поставленої мети в кваліфікаційній роботі необхідно вирішити наступні **завдання**:

- дослідити предметну область з точки зору системного аналізу;
- проаналізувати сучасні методи інтелектуального аналізу для обробки медичних даних;
- ідентифікувати важливі прогностичні фактори ризику серцево-судинних захворювань;
- розробити модель прогнозування серцево-судинних захворювань на основі виявлених факторів;
- перевірити ефективність розробленої моделі на реальних медичних даних;
- оцінити вплив розробленої моделі на процес діагностики, лікування та профілактики серцево-судинних захворювань.

В кваліфікаційній роботі використовувалися загальнонаукові та спеціальні методи наукового дослідження такі, як системний аналіз, синтез, порівняння, статистичний аналіз, класифікація, кореляція, аналіз виживання тощо.

Теоретична значущість проведеного дослідження полягає в тому, що було узагальнено наукові знання щодо питань використання методів інтелектуального аналізу для покращення прогнозування серцево-судинних захворювань, дослідження може стати основою для подальших наукових робіт у цій області.

Практичне значення одержаних результатів полягає у тому, що вони можуть бути використані в медичній практиці для покращення прогнозування серцево-судинних захворювань. Це, в свою чергу, може допомогти призначити вчасне та ефективне лікування, знизити втрати на медичне обслуговування та оптимізувати використання ресурсів.

Апробація результатів дослідження. Основні положення та результати дослідження доповідалися на V International Scientific and Theoretical Conference «The process and dynamics of the scientific path» (Афіни, Грецька Республіка, 2024 р.)

Структура кваліфікаційної роботи відповідає меті та завданням дослідження та складається зі вступу, основної частини з трьох розділів, висновків, списку

використаних джерел та додатків. Робота викладена на 67 сторінках, містить 26 рисунків, 1 таблицю. Список використаних джерел включає 55 джерел.

ВИСНОВКИ

Кваліфікаційна робота була присвячена методам інтелектуального аналізу даних для покращення прогнозування серцево-судинних захворювань.

В ході дослідження були отримані наступні результати:

1. Виконано системний аналіз методів та моделей прогнозування серцево-судинних захворювань. Зазначено, що ключем до прогнозування ризику ССЗ є своєчасне прогнозування, комплексна профілактика, та виявлення факторів ризику. Показано, що алгоритми оцінки ризиків ССЗ, які використовуються зараз, не відповідають сучасним вимогам, що потребує пошуку нових методів та підходів. Визначено, що одним із перспективних напрямів є створення прогнозних та діагностичних моделей з використанням методів ІАД.

2. Проаналізовано основні методи інтелектуального аналізу обробки медичних даних. Зазначено, що метою ІАД є створення моделі, яка може ефективно інтерпретувати та використовувати дані. Процес побудови моделі ІАД можна представити як послідовність наступних базових кроків: збір даних про стан здоров'я пацієнтів, визначення проблеми та формулювання завдання, попередня обробка даних, вивчення даних, створення моделі, дослідження та перевірка моделі, розгортання та оновлення моделі. Наведено процес побудови моделі інтелектуального аналізу медичних даних та дана оцінка її продуктивності.

3. Survival Analysis, аналіз виживання або аналіз настання подій – це метод, за допомогою якого протягом визначеного періоду часу вивчають закономірності появи певного результату у представників спостережуваної вибірки. Під методами оцінки виживання розуміється вивчення закономірності появи очікуваної події у представників спостережуваної вибірки у часі. Визначено поняття цензурованих даних, які характеризують процес, який на момент останнього обстеження не закінчується досліджуваним результатом. Доведено поняття «функції виживання» як ймовірності того, що час смерті настане пізніше за деякий заданий час.

4. Виявлено фактори ризику серцево-судинних захворювань за допомогою метода Каплана-Майєра. Представлено графічне уявлення методу Каплана-Майєра, котре полягає у побудові кривої виживання, яка відображає пропорцію пацієнтів, у яких очікувана подія не відбулася до певного моменту часу. Для оцінки кривих Каплана-Майєра використано логарифмічний ранговий тест, який встановлює ефективність нового лікування порівняно з контрольним.

5. Для більш точного аналізу виживання та дослідження зв'язку між часом виживання пацієнтів та встановленими факторами ризику, за допомогою метода аналізу виживання, кривих Каплана-Майєра та регресійної моделі пропорційних ризиків Коксу, розроблено модель прогнозування серцево-судинних захворювань.

6. Доведено, що дана група методів аналізу виживання визначає групи пацієнтів, які перебувають під великим ризиком. Встановлено статистично значущі фактори ризику, які дають можливість медичним працівникам зосередити свої зусилля на пацієнтах групи ризику та забезпечують їх інтенсивним лікуванням, дозволяють знизити втрати на медичне обслуговування та оптимізувати використання ресурсів. Відмічено, що ефективність лікування серцево-судинних захворювань значною мірою залежить від точності прогнозу, своєчасної діагностики, комплексної профілактики та виявлення факторів ризику.