**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**Чорноморський національний університет**
**імені Петра Могили**
**Факультет комп'ютерних наук**
**Кафедра інтелектуальних інформаційних систем**

**ДОПУЩЕНО ДО ЗАХИСТУ**

Завідувач кафедри інтелектуальних
інформаційних систем, д-р техн. наук, проф.,
_____Ю. П. Кондратенко
«____» _____2022 р.

БАКАЛАВРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

# ІНФОРМАЦІЙНА СИСТЕМА КЛАСИФІКАЦІЇ ДЛЯ ОЦІНЮВАННЯ СТАНУ ЗДОРОВ'Я ЛЮДИНИ

Спеціальність 122 «Комп'ютерні науки»

122 – БКР – 402.21810206

Виконав: студент 4 курсу, групи 402
_____ С.Ю. Воздіцький
« 21 » червня 2022 р.

Керівник: кандидат техн. наук, доцент,
доцент кафедри ІІС
_____ І.О. Калініна
« 21 » червня 2022 р.

**Миколаїв – 2022**

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**Чорноморський національний університет ім. Петра Могили**
**Факультет комп'ютерних наук**
**Кафедра інтелектуальних інформаційних систем**

Рівень вищої освіти **бакалавр**
Спеціальність **12 «Комп'ютерні науки»**
*(шифр і назва)*
Галузь знань **12 «Інформаційні технології»**
*(шифр і назва)*

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних інформаційних систем, д-р техн. наук, проф._____Ю.П. Кондратенко
«___»_____2022 р.

**ЗАВДАННЯ**
**на виконання кваліфікаційної роботи**

Видано студенту групи 402 факультету комп'ютерних наук
_____Воздіцькому Сергію Юрійовичу_____
*(прізвище, ім'я, по батькові)*
1. Тема кваліфікаційної роботи
«Інформаційна система класифікації для оцінювання стану здоров'я людини»

Керівник роботи Калініна Ірина Олександрівна кандидат техн. наук, доцент, доцент кафедри ІІС
*(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)*
Затв. наказом Ректора ЧНУ ім. Петра Могили від «07» грудня 2021 р. № 318

2. Строк представлення кваліфікаційної роботи студентом «__» червня 2022 р.

3. Вхідні (початкові) дані до роботи: загальна інформація про інформаційні системи, їх види, особливості медичних інформаційних систем, методи класифікації, моделі класифікаторів, відомості про інтелектуальний аналіз даних.

Очікуваний результат роботи: створення інформаційної системи для оцінювання стану здоров'я людини, а саме діагностування ішемічної хвороби серця шляхом вивчення даних про пацієнтів, їх результати аналізів, симптоми, шкідливі звички тощо.

4. Перелік питань, що підлягають розробці (зміст пояснювальної записки )
1. Аналіз актуальності поставленої задачі про розробку інформаційної системи.
2. Вивчення загальної теорії.
3. Розгляд засобів програмної реалізації.
4. Програмна реалізація інформаційної системи класифікації для діагностування пацієнтів.
5. Тестування та отримання результатів роботи інформаційної системи.

5. Перелік графічних матеріалів 11 таблиць, 9 рисунків, презентація.

6. Завдання до спеціальної частини «Охорона праці з правил дотримування показників мікроклімату у робочих приміщеннях»

7. Консультанти:

| Розділ | Прізвище, ініціали та посада консультанта | Підпис |
|---|---|---|
| Охорона праці | Алексєєва А.О. | |
| | | |

Керівник роботи кандидат техн. наук, доцент, доцент кафедри ІІС Калініна І.О.

*(наук. ступінь, вчене звання, прізвище та ініціали)*

*(підпис)*

Завдання прийнято до виконання Воздіцький С.Ю.

*(прізвище та ініціали)*

*(підпис)*

Дата видачі завдання « 25 » листопада 2021 р.

**КАЛЕНДАРНИЙ ПЛАН**
**виконання бакалаврської кваліфікаційної роботи**

Тема: Інформаційна система класифікації для оцінювання стану здоров'я людини

| № | Найменування роботи | Початок | Закінчення | Примітки |
|---|---|---|---|---|
| 1 | Подання заяви на затвердження теми та керівників БКР | 18.11.2021 | 19.11.2021 | Виконано |
| 2 | Отримання завдання на виконання БКР | 24.12.2021 | 24.12.2021 | Виконано |
| 3 | Складання календарного плану роботи на весь період виконання БКР | 24.12.2021 | 24.12.2021 | Виконано |
| 4 | Отримання завдання на переддипломну практику | 20.05.2022 | 20.05.2022 | Виконано |
| 5 | Проходження переддипломної практики, збір та аналіз матеріалів до БКР | 23.05.2022 | 01.06.2022 | Виконано |
| 6 | Розробка звіту з переддипломної практики | 02.06.2022 | 04.06.2022 | Виконано |
| 7 | Виконання БКР: аналіз сучасного стану задачі запису до електронної черги, огляд існуючих аналогів та технологій, розробка ПЗ | 10.02.2022 | 28.05.2022 | Виконано |
| 8 | Попередній захист БКР на засіданні комісії кафедри | 30.05.2022 | 30.05.2022 | Виконано |
| 9 | Доробка та остаточне оформлення БКР | 01.06.2022 | 03.06.2022 | Виконано |
| 10 | Подання БКР рецензенту | 04.06.2022 | 04.06.2022 | Виконано |
| 11 | Подання БКР, її електронної копії та інших документів (відгуку, рецензії) до захисту | 05.06.2022 | 05.06.2022 | Виконано |
| 12 | Захист БКР перед екзаменаційною комісією (ЕК) | 27.06.2022 | 29.06.2022 | Виконано |

Розробив студент Воздіцький С. Ю. _____
*(прізвище та ініціали)* *(підпис)*

Керівник роботи к. т. н., доцент, Калініна І.О. _____
*(наук. ступінь, вчене звання, прізвище та ініціали)* *(підпис)*

«12» грудня 2021 р.

**ВІДГУК**
**на бакалаврську кваліфікаційну роботу студента групи 402**
**ЧНУ імені Петра Могили**
**Воздіцького Сергія Юрійовича**
**«Інформаційна система класифікації для оцінювання стану здоров'я людини»**

У сучасному світі не можливо уявити теперішнє життя без інформаційних технологій, а саме без інформаційних систем, які оточують нас майже всюди. Тому розвиток ІТ-технологій у напрямку медицини є не менш важливим, ніж у інших галузях. Саме цьому, темою бакалаврської кваліфікаційної роботи Воздіцького Сергія Юрійовича є розробка медичної інформаційної система класифікації для оцінювання стану здоров'я людини та діагностування ішемічної хвороби серця.

Робота присвячена інтелектуальному аналізу даних та вирішення задачі класифікації на базі тих даних, які представлені у загальнодоступному наборі даних про пацієнтів та їх атрибути. Для реалізації поставленої задачі були обрані сучасні інструменти та технології, які дозволяють найбільш правильно класифікувати ті чи інші дані та робити висновки.

У дипломній роботі дуже детально розглянуто усі поняття, які стосуються теми, проведено аналіз переваг та недоліків, обрано найкращі варіанти вирішення та реалізовано найбільш точну програму.

Робота складається з трьох розділів, у яких послідовно розглянуто усі поняття, більш детально пояснено складні моменти. Матеріал поданий професійно.

Бакалаврська робота Воздіцького Сергія Юрійовича проведена на високому рівні. На мою думку, слід допустити роботу Воздіцького С.Ю. до захисту та присвоїти йому освітню кваліфікацію «бакалавр комп'ютерних наук» в галузі знань 12 «Інформаційні технології» за спеціальністю 122 «Комп'ютерні науки».

Керівник
кандидат техн. наук,
доцент, доцент кафедри ІІС _____ І.О. Калініна

**АНОТАЦІЯ**

**бакалаврської кваліфікаційної роботи студента групи 402 ЧНУ ім. Петра Могили**

**Воздіцького Сергія Юрійовича**

**Тема: «Інформаційна система класифікації для оцінювання стану здоров'я людини»**


У роботі розібрано та досліджено найпоширеніші методи інтелектуального аналізу даних та рішення задачі класифікації. Вивчено різні методи класифікації. Робота обраних методів була розглянута на практичній задачі, а саме оцінка стану здоров'я.

Об'єктом дослідження стали медичні показники людини, які здобуваються методами опитування, аналізів, досліджень.

Предметом дослідження стали математичні моделі класифікації.

Робота складається з фахового розділу і спеціальної частини з охорони праці. Пояснювальна записка складається зі вступу, чотирьох розділів, висновків та додатків.

Бакалаврська кваліфікаційна робота містить 55 с., 9 рис., 11 табл., 1 додаток, 23 джерела.



Ключові слова: класифікація, інформаційні системи, інтелектуальний аналіз даних, методи, метрики, набори даних.

**ABSTRACT**

**Bachelor's qualification work of the student of 402 group of Petro Mohyla Black Sea National University Vozditskogo Sergey Yurievicha**

**Title: “Classification information system for assessing human health”**

The paper analyzes and investigates the most common methods of data mining and classification solutions. Different classification methods have been studied. The work of the selected methods was considered in the practical task, namely the assessment of health.

The object of research was medical indicators of a person, which are obtained by methods of survey, analysis, research.

Mathematical models of classification became the subject of research.

The work consists of a professional section and a special section on labor protection. The explanatory note consists of an introduction, four chapters, conclusions and appendices.

Bachelor's thesis contains 55 pages., 9 fig., 11 table., 1 app., 23 sources.

Key words: classification, information systems, intellectual data analysis, methods, metrics, data sets.

# ЗМІСТ

# ПЕРЕЛІК СКОРОЧЕНЬ

БД – база даних

ЕОМ – електронна обчислювальна машина

ІАМД – інтелектуальний аналіз медичних даних

ІС – інформаційна система

МІС – медична інформаційна система

МСА – математично-статистичний аналіз

ДС – дата сет

DB – data base

DS – data set

# Пояснювальна записка

**до кваліфікаційної роботи**

на тему:

## «ІНФОРМАЦІЙНА СИСТЕМА КЛАСИФІКАЦІЇ ДЛЯ ОЦІНЮВАННЯ СТАНУ ЗДОРОВ'Я ЛЮДИНИ»

Спеціальність 122 «Комп'ютерні науки»

**122 – БКР – 402.21810206**


Виконав: студент 4-го курсу, групи 402
_____С.Ю. Воздіцький_____
*(підпис, ініціали та прізвище)*

« 21 » червня 2022 р.

Керівник: к. т. н., доцент.каф.ІІС_____
*(наук. ступінь, вчене звання)*
_____І.О.Калініна_____
*(підпис, ініціали та прізвище)*

« 21 » червня 2022 р.


**Миколаїв – 2022**

# ВСТУП

Поняття інформаційних систем вперше були впроваджені у п’ятдесятих роках минулого століття. Призначення таких систем були обробки рахунків та розрахунки заробітних плат, а реалізація проводилася на електромеханічних бухгалтерських лічильних машинах. Завдяки цьому, це призводило до деяких скорочень часу на підготовку паперових документів та витрат на усі ці дії.

Згодом, технології рухались вперед, усе більші об’єми інформації потребували більш розвинених інформаційних систем. Шістдесяті роки відомі зміною ставлення до ІС. Вся інформація та дані, отримані з них, почали застосовуватися для періодичної звітності за багатьма параметрами. Звісно, для цього різним корпораціям та організаціям було потрібно комп’ютерне обладнання широкого призначення, яке буде здатне обслуговувати велику кількість задач.

На межі сімдесятих та восьмидесятих років інформаційні системи починають використовувати як засіб управлінського контролю, що прискорює та підтримую процес прийняття різноманітних рішень. Концепція використання інформаційних систем знову починає змінюватися до кінця восьмидесятих років минулого століття. Тоді вони вже стають стратегічним джерелом інформації та використовуються на всіх рівнях організації будь-якого профілю. ІС того періоду, тепер у стані надавати вчасно потрібну інформацію, допомагають різним корпораціям та організаціям досягти успіху

у своїй діяльності, створювати нові товари та послуги, знаходити нові ринки збуту, забезпечувати собі гідних партнерів та організовувати випуск продукції за низькою ціною тощо.

Інформаційні системи є актуальними і наразі. Сучасне життя просто неможливо уявити без усіх цих систем, які сьогодні оточують нас майже у кожній сфері життя. Тому для подальшого розвитку людства та благ, які надають нам інформаційні технології, важливо вміти впізнавати, розрізняти та створювати інформаційні системи.

Сучасне розуміння ІС передбачає використання як основний технічний засіб переробки інформації персонального комп'ютера. Крім того, технічне втілення інформаційної системи саме по собі нічого не означає, якщо не враховано роль людини, для якої призначена вироблена інформація і без якої неможливе її отримання та уявлення.

Необхідно розуміти різницю між комп'ютерами та інформаційними системами. Комп'ютери, оснащені спеціалізованими програмними засобами, є технічною базою та інструментом ІС. Інформаційна система не може існувати без персоналу, що взаємодіє з комп'ютерами та телекомунікаціями.

В даний момент часу панує думка про інформаційну систему як про систему, реалізовану за допомогою комп'ютерних технологій. Хоча у загальному трактуванні інформаційну систему можна розуміти й у некомп'ютерному варіанті.

Сама ідея інформаційних систем та деякі принципи їх організації виникли задовго до появи електронних обчислювальних машин. Бібліотеки, архіви, адресні бюро, телефонні довідники – це інформаційні системи. Проте комп'ютеризація у декілька разів підвищила ефективність інформаційних систем та розширила сфери їх застосування.

**Метою даної роботи** є створення інформаційної системи класифікації для оцінювання стану здоров’я людини. На сьогоднішній день, коли питання здоров’я стало ребром для людства, що в першу чергу відобразилося на житті кожної людини та на економіках країн світу, дуже важливо застосовуючи увесь потенціал і накопичені знання та вміння людства, долати глобальні проблеми завдяки й інформаційним технологіям. Завдяки такій ІС, можна за медичними показниками людини оцінювати та прогнозувати стан її здоров’я.

**Об’єктом дослідження** виступає процес класифікації медичних показників пацієнта, наприклад симптоми, аналізи, результати обстежень, та їх значення для успішного діагностування захворювання.

**Предмет дослідження** – моделі методи бінарної класифікації на основі структурованого набору даних.

**Методологія і методи досліджень** – для досягнення визначених задач були використані технології програмування, а також вивчення медичної літератури для правильної оцінки стану здоров’я людини.

**Апробація результатів** дипломної роботи.

**Практична значимість** даної роботи у тому, що завдяки цій розробленій інформаційній системі можна проводити оцінювання стану здоров’я пацієнта у медичних закладах, або у приватному порядку. Така ІС зможе значно полегшити роботу медичних працівників, наприклад давати певний прогноз про стан здоров’я обстежуваного пацієнта, який можна використовувати при подальшому курсу лікування.

**Публікації.** За результатами бакалаврської роботи опубліковані тези доповіді.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ СФЕРИ ІС ОЦІНКИ СТАНУ ЗДОРОВ'Я ТА ПОСТАНОВКА ЗАДАЧІ

## 1.1 Актуальність та проблеми предметної області

### 1.1.1 Поняття інформаційної системи

Під системою розуміється будь-який об'єкт, який може одночасно розглядатися і як щось єдине ціле, і як об'єднана для досягнення поставлених задач сукупність різноманітних елементів. Системи насправді, сильно відрізняються одна від одної між собою: по складу та головним цілям.

Інформаційна система – це взаємопов'язана сукупність засобів, методів та персоналу, які використовуються для зберігання, обробки, пошуку та видачі інформації заради досягнення поставленого завдання.

ІС можна представити як деяке сховище інформації, забезпечене процедурами введення цієї інформації, пошуку, розміщення та видачі її. Наявність таких процедур – головна особливість інформаційних систем, що головним чином відрізняють їх від простих накопичень інформаційних матеріалів. Можна привести декілька прикладів для більш ясного розуміння цієї різниці. Інформаційною системою, наприклад особиста бібліотека, де орієнтуватися у ній може лише власник цієї бібліотеки, вважатися не може. А у публічних бібліотеках порядок розміщення книг завжди чітко визначений. Тому завдяки цьому пошук та видача цих книг, а також розміщення нових надходжень є стандартними процедурами, дуже близькими до алгоритмів.

Важливо також розуміти якими властивостями може визначатися ІС:

a) ІС є динамічною та вона розвивається;
b) при розробці та побудуванні ІС треба використовувати системний підхід;
c) ІС необхідно сприймати як систему обробки даних подібну до «людини-комп’ютера»;
d) люба ІС може бути піддана аналізу, побудована та керована на основі загальних принципів побудови систем;
e) вихідною продукцією ІС є інформація, на основі якої підтримуються та приймаються рішення;

Процеси, які забезпечують роботу ІС будь-якого призначення, умовно можна поділити на такі складові:

a) введення інформації із зовнішніх або внутрішніх джерел даних;
b) обробка вхідних даних та їх подання у зручному вигляді;
c) виведення інформації для подання споживачам або передачі до іншої системи;
d) зворотній зв’язок – це інформація, перероблена людьми цієї організації для корекції вхідних даних;

Функціювання інформаційних систем полягає у обслуговуванні двох зустрічних потоків даних, наприклад: введення нової інформації та видачі поточної інформації за запитами. Так як головне завдання ІС це

обслуговування користувачів цієї системи, вона повинна бути налаштована так, щоб відповідь на будь-який запит видавався швидко та вона була у більшій мірі повною. Ці вимоги забезпечуються наявністю стандартних процедур пошуку інформації та тим, що ці системи розташовані у порядку.

Як вже було зазначено вище, масова комп’ютеризація у 80-90-х роках розширила сфери використання інформаційних систем та підвищила їх ефективність. По-перше, відбулось різке зростання швидкості усіх видів обробки даних: їх пошук та розміщення всередині самої електронної обчислювальної машини, видачі на екран або друк інформації, передачі даних засобами космічного та електронного зв’язку у будь-яку точку земної кулі. Для деяких видів інформаційних систем, саме покажчик швидкості передачі та введення даних відіграє вирішальну роль. Можна навести такі приклади: автоматизовані системи продажу квитків або багатотермінальні системи електронної торгівлі цінними паперами, де тільки висока швидкість обробки та передачі даних може виключити продажу акцій, які декілька хвилин тому були продані з іншого терміналу. По-друге, у велику кількість разів збільшилися можливості зберігання великих обсягів інформації: як за рахунок того, що паперові носії даних не такі компактні як машинні електронні, які можуть зберігати інформацію в сотні і тисячі разів більше так і за рахунок того, що тільки при високих швидкостях ЕОМ можна проводити пошук у таких обсягах за прийнятний час.

По-третє, завдяки використанню електронного зв'язку та мереж ЕОМ втратила значення відстань між ІС, джерелами інформації та її клієнтами. Досить мати термінал, тобто персональну ЕОМ або інший пристрій, що дозволяє запитувати та отримувати потрібні дані та з'єднаний із інформаційною системою каналами зв'язку.

Не слід вважати, що висока ефективність сучасних інформаційних систем автоматично досягається застосуванням сучасних технічних засобів та технологій. Щоб максимально використовувати їх величезні можливості, потрібно добре опрацювати мовні, алгоритмічні і структурні питання, тобто розробити структури даних, алгоритми обробки інформації та мови спілкування із системою.

### 1.1.2 Аналіз різновидів інформаційних систем

Вид інформаційної системи залежить від того, чиї інтереси вона обслуговує та на якому рівні управління. За характером подання та логічної організації збереженої інформації, ІС поділяються на:

a) фактографічні ІС;

b) документальні ІС;

c) геоінформаційні ІС.

Фактографічні інформаційні системи зберігають дані у вигляді множини екземплярів одного або декількох типів структурних елементів, або по-іншому кажучи інформаційних об'єктів. Кожен з таких екземплярів або

деяка їхня сукупність відображають відомості за будь-яким фактом, подією окремо від усіх інших відомостей та фактів. Структура кожного типу інформаційного об'єкта складається з кінцевого набору реквізитів, що відображають основні аспекти та характеристики об'єктів даної предметної галузі. Комплектування інформаційної бази у фактографічних інформаційних системах включає, зазвичай, обов'язковий процес структуризації вхідної інформації. Фактографічні інформаційні системи передбачають задоволення інформаційних потреб безпосередньо, тобто шляхом подання споживачам самих відомостей, наприклад: даних, фактів, концепцій.

У документальних або документованих ІС одиничним елементом інформації є нерозділений та нерозчленований на більш дрібні елементи документ або інформація. Під час введення, яке називається вхідний документ, зазвичай, ця інформація не структурується, чи структурується в обмеженому вигляді. Для документа, що вводиться, можуть встановлюватися деякі формалізовані позиції, такі як дата виготовлення, виконавець, тематика тощо. Деякі види документальних ІС забезпечують встановлення логічного зв'язку документів, що вводяться, наприклад: взаємні посилання за будь-якими критеріями, підпорядкованість за змістом і т.д. Визначення та встановлення такого взаємозв'язку є складною багатокритеріальною та багатоаспектною аналітичною задачею, яка не може бути формалізована у повній мірі.

У геоінформаційних системах або ГІС, набір даних організований як окремі інформаційні об'єкти з певним набором реквізитів, прив'язаних до загальної електронної топографічної основі, наприклад електронної карти. ГІС застосовуються для інформаційного забезпечення у тих предметних галузях, у структура інформаційних об'єктів та процесів яких має просторово-географічний компонент (господарство, маршрути транспорту, підприємства тощо).

Також інформаційні системи можна розрізняти за деякими іншими критеріями, наприклад один з таких - за функціональною ознакою. Так, функціональна ознака визначає призначення підсистеми, а також її основні цілі, завдання та функції. Нижче на рис. 1.1 представлено класифікацію ІС за характеристикою їх функціональних підсистем у вигляді схеми.


Рис. 1.1. - Класифікація ІС за функціональною ознакою

У господарській практиці виробничих та комерційних об'єктів типовими видами діяльності, що визначають функціональну ознаку

класифікації інформаційних систем, є фінансова, маркетингова, кадрова та виробнича діяльність.

Важливо класифікувати інформаційні системи й за рівнями керування. Наприклад, можна виділити такі:

d) ІС оперативного або операційного рівня.

До таких інформаційних систем можна віднести бухгалтерські, банківських депозитів, реєстрації квитків, виплати зарплат, обробки замовлень тощо. ІС оперативного або операційного рівня підтримує фахівців-виконавців, обробляючи дані про угоди та події, наприклад рахунки, накладні, зарплата, кредити, потік сировини та матеріалів тощо. Призначення інформаційної системи на цьому рівні — це відстежувати потік угод та відповідати на запити про поточний стан у фірмі, що відповідає оперативному керуванню. Щоб виконувати таке завдання, ІС має бути легкодоступною, безперервно чинною та надавати точну інформацію про стан.

Завдання, мета та джерела інформації на операційному рівні заздалегідь визначені та високо впорядковані. Рішення запрограмоване відповідно до запитуваного алгоритму дій.

Інформаційна система оперативного рівня є сполучною ланкою між фірмою та зовнішнім середовищем. Наприклад, якщо система працює не дуже добре, то організація або отримує інформацію з зовнішній джерел, або не видає інформацію взагалі. Але з іншого боку, система — це основний постачальник

даних інших типів інформаційних систем у організації чи корпорації, так як містить і архівну, і оперативну інформацію.

- Інформаційні системи тактичного рівня;

Основними функціями таких ІС є забезпечення доступу до архівної інформації, складання періодичних звітів за певний проміжок часу, порівняння поточних показників з минулими. До таких систем можна віднести системи підтримки прийняття рішень (СППР). Вони працюють з частково структурованими задачами, результати яких важко спрогнозувати заздалегідь, бо вони мають потужніший аналітичний апарат із декількома моделями. Дані у такі системи постачаються з оперативних та управлінських інформаційних систем. Перелік посад, які використовують ці системи у певних організаціях дуже великий, так як це люди яким необхідно приймати рішення. Це можуть бути менеджери, спеціалісти, аналітики тощо. Наприклад, їх рекомендації можуть знадобитися при прийнятті рішення купувати або взяти умовне майно в оренду. Дуже важливо виділяти характеристику СППР, цим самим можна більш точніше визначити доречність їх використання. Наприклад, системи прийняття рішень забезпечують вирішення проблем, появу та поведінку яких важко спрогнозувати; вони дозволяють легко та швидко замінювати постановки задач та вхідні дані; системи відрізняються гнучкістю та легко адаптуються до зміни умов за дуже короткий період часу, наприклад – кілька разів на день;

вони оснащенні складним інструментарієм моделювання та аналізу; в решті решт СППР мають технології, які максимально є орієнтованими на користувача;

- Інформаційні системи спеціалістів;

Інформаційні системи такого рівня сприяють роботі фахівців, які працюють з даними, підвищують продуктивність та продуктивність роботи проектувальників або інженерів. Завдання подібних ІС - інтеграція нових відомостей та даних в організацію або підприємство та допомогу в обробці паперових документів. В залежності від того, як індустріальне суспільство перетворюється у інформаційне, продуктивність та ефективність все більше буде залежати від рівня розвитку цих систем. Такі системи, наприклад у вигляді офісних систем та робочих станцій, найбільше затребувані та найшвидше розвиваються. У якості прикладу можна привести ІС офісної автоматизації. Такі системи внаслідок своєї багатопрофільності та простоти у використанні, активно застосовуються працівниками будь якого рівня. Частіше їх використовують такі працівники: клерки, секретарі, бухгалтери і т.д. Головна мета – це підвищити ефективність їх роботи, спростити канцелярську складову та обробка інформації та даних. Системи офісної автоматизації пов'язують працівників цього інформаційного середовища у різних місцях та сприяють підтримці зв'язка з клієнтами: покупцями, замовниками та іншими організаціями. Діяльність таких систем переважно

охоплює управління комунікацією, документацією, складанням розкладів тощо. Можна виділити такий перелік функцій цих систем:

a) архівація документів;

b) виробництво друкованих матеріалів;

c) електронні записники для ведення справ;

d) електронна пошта;

e) обробки великих обсягів тексту на ЕОМ;

f) телеконференції;

ІС обробки знань як і експертні системи включають у собі знання, необхідні різним спеціалістам, таким як вченим, юристам, інженерам, при розробці чи створенні нової продукції. Наприклад, забезпечення високого рівня технічних розробок зможуть спеціалізовані робочі станції з того чи іншого проектування.

- Стратегічні ІС.

Успіх той чи іншої організації найчастіше визначається тим, яку стратегію для неї затверджено. Стратегія – мається на увазі набір методів та способів рішення довгострокових перспективних завдань. У цьому криються такі поняття як стратегічний засіб, метод, система. Перехід до ринкових відносин у сьогодення, звернув велику увагу до питання стратегії розвитку та поведінки організації та й це сприяло докорінній зміні у поглядах стосовно інформаційних систем. Звісно вони стали розцінюватися як стратегічно важливі та які можуть впливати на зміну цілей тої чи іншої організації або

фірми, її завдань, продукції, методів, що дозволяє випереджати конкурентів. Тому треба виділити основне поняття, що таке стратегічна інформаційна система. Стратегічна ІС – це комп’ютерна інформаційна система, що забезпечує та підтримує прийняття рішень стосовно реалізації перспективних стратегічних цілей розвитку фірми чи іншої організації. Можна виділити окремі ситуації, коли нова якість ІС спричиняла зміни структур, профілів організацій сприяючи їх процвітанню. Але при цьому з’являється риск появи небажаної психологічної обстановки у середовищі, пов'язане з автоматизацією деяких видів робіт та функцій, так як це може поставити деяку частину працюючих у погане становище.

Окремо виділяють й інші класифікації ІС:

a) за рівнем автоматизації;
    1) ручні ІС;
    2) автоматичні ІС;
    3) автоматизовані ІС;
b) за характером використання інформації;
    1) радячі ІС;
    2) керуючі ІС;
    3) інформаційно-пошукові ІС;
    4) інформаційно-вирішальні ІС;
c) у сфері застосування;
    1) інтегровані (корпоративні) ІС;

2) ІС автоматизованого проектування;

3) ІС управління технологічними процесами;

4) ІС організаційного управління;

d) за способом організації;

1) на основі інтернет-технологій;

2) на основі систем з урахуванням архітектури файл-сервер;

3) на основі архітектури клієнт-сервер;

4) на основі багаторівневої архітектури;

## 1.2 Опис медичних інформаційних систем

### 1.2.1 Поняття медичної діагностики

Сучасна медицина базується у переважній більшості на методах діагностування. Діагностика – наука про принципи та способи ідентифікування та створення нових методів розпізнавання існуючих хвороб. Мета такої науки – поставлення правильного діагнозу та вибір правильного ефективного курсу лікування яке зможе покращити стан здоров’я захворюваних, або призвести до повного їх одужання.

Основне джерело інформації для діагностування базуються на різноманітних аналізах та дослідженню показників здоров’я людини. Нижче приведені базові методи дослідження:

1. аналізи крові – усім відомий базовий покажчик. Основуючись на даних з такого аналізу, лікар може визначити чи є в організмі

наприклад запальний процес, склад речовин у крові та загалом визначити стан захворювання;

2. сучасні прилади дозволяють наочно досліджувати травний тракт пацієнта;
3. завдяки ЕКГ яке дозволяє проводити моніторинг роботи серця, можна проводити кардіографічні аналізи;
4. такі технології як МРТ, КТ, УЗД – дозволяє оцінити стан внутрішніх органів людини на приклад патологій, травм тощо;

### 1.2.2 Обробка медичних даних

З кожним роком, прогрес в медицині не стоїть на одному місці. Збільшується кількість хвороб, методів дослідження та лікування, пацієнтів тощо. Тому одна з найважливіших проблем в процесі роботи є самі дані, їх об’єм та способи обробки таких великих обсягів.

Як було зазначено у минулому підрозділі, з початку вісімдесятих та дев’яностих років, було запроваджено масове інтегрування медичних інформаційних систем (МІС) – систем, головним завданням яких була автоматизація всіх основних процесів, які пов’язані із роботою медичних закладів усіх спеціалізацій. Такі автоматизовані ІС дозволяють ефективно й швидко налагодити документообіг, пропрацювати процес взаємодії з пацієнтами та контролювати всі адміністративні та фінансові питання.

Треба чітко розуміти, яким стрімким є щорічний приріст даних в таких системах. Можна привести приклад: у 2012 році електронні дані у сфері охорони здоров’я оцінювались експертами у 500 петабайт (1 петабайт – 1024 терабайт), то за прогнозами на 2022 рік, така цифра вже буде становити приблизно 25 000 петабайт. Але можливості, які надають МІС, варті того, щоб розвивати надалі технології зберігання та обробки даних.

### 1.2.3 Переваги медичних інформаційних систем

Для того, щоб вкладати ресурси у розвиток технології створення та підтримування таких систем, важливо виділяти їх переваги. Перш за все, інформаційні системи в медицині дозволяють керувати великі маси даних про пацієнтів та їх результати. Тепер вся інформація, яка зберігається в МІС, доступна будь-де у будь-який час. Тим самим уніфікується підхід до людей які лікуються, а вся інша медична документація може створюватись за певним зразком.

Далі, відбувається злиття даних та їх звітність, що дозволяє розробляти електронні структури для будь-яких структурних одиниць, наприклад лікарень або окремих кабінетів, та об’єднувати їх у єдину електронну систему.

Внесення даних у медичні інформаційні системи роблять такі дані доступними для обробки та аналізу. Наприклад, дата сет, який використовується у даній роботі, є доступним завдяки існування таких

інформаційних систем. На рисунку 1.2 показана спрощена схема регулювання даних у МІС.


Рис. 1.2 – схема регулювання даних у МІС

Можна коротко підвести висновки щодо переваг використання медичних інформаційних систем. Звісно, переваги будуть покривати усі недоліки, якщо така система була обрана та розроблена вдало. Тоді яку користь надає МІС для медичних закладів? Приклади приведено нижче:

- позбавляє від заповнення паперів;
- зниження впливу людського фактору;
- підвищення працездатності та якості обслуговування;
- повсюдний доступ до інформації;

## 1.3 Методи обробки даних в медицині

Обробка даних або інтелектуальний аналіз медичних даних (ІАМД) є надзвичайно важливим у сучасній доказовій медицині. Його методи є інструментом як функціоналом інтелектуальних систем підтримки прийняття медичних рішень, так і безпосереднього оброблення результатів клінічного дослідження. До прикладу: одним з методів ІАМД є математико-статистичний аналіз (МСА) результатів клінічних досліджень. Такі дослідження дозволяють обчислити статистичні параметри наборів даних, оцінити параметри генеральної сукупності, виявити та встановити закономірності та причинно-наслідкові зв’язки між даними. А результати таких інтелектуальних опрацювань є підґрунтям для прийняття чи відхилення тих чи інших рішень.

Не беручи до уваги існування великої кількості наукових матеріалів на тему методології МСА, проблема вибору методів математично-статистичного аналізу які є адекватними завданням дослідження, на сьогодні є актуальним.

### 1.3.1 Перевірка відповідності розподілу значень

Звичайно перевірка відповідності розподілу значень кількісних та якісних ознак закону нормального розподілу допускає, що цей розподіл в генеральній сукупності визначається за вибіркою. Такі перевірки можна виконати різними способами.

Наприклад, є якісний спосіб побудови гістограми розподілу значень ознаки та візуальна її оцінка – наскільки вона близька до нормального розподілу. Тепер треба дати визначення поняттю гістограма – це стовпчаста діаграма, яка відображає частотні дані.

Також можна звернути увагу на кількісний спосіб обчислення дескриптивних статистик та їх порівняння. Такі статистики обчислюються за допомогою відповідних інструментальних засобів пакетів прикладних програм математично-статистичного аналізу або завдяки бібліотечних статистичних функцій табличних процесорів. Інший кількісний спосіб включає у себе оцінку симетричності розподілу ознаки з тільки додатними значеннями, але при цьому неможливо оцінити ексцес.

Існує і надійний спосіб – це перевірка статистичних гіпотез про вид розподілу. Формулюються дві гіпотези: нульова – про те, що розподіл ознаки в генеральній сукупності відповідає закону НР – нормального розподілу; та альтернативна гіпотеза – про те, що розподіл ознаки в генеральній сукупності не відповідає закону НР. Важливо підкреслити, що перевірка статистичних гіпотез про вид розподілу значень ознаки відбувається на підставі статистичних критеріїв узгодженості. Прикладами таких узгодженостей можуть бути: Шапіро-Вілка, Д’Агостіно, Крамера-Мізеса, Колмогорова-Смірнова, Андерсона-Дарлінга.

Отже, якщо розподіл кількісних значень ознаки нормальний або близький до нього, то математично-статистичний аналіз, а саме порівняння

груп за цією ознакою, використовується завдяки параметричним методам. А якщо розподіл відмінний від нормального або є категоріальним чи якісним, то для порівняння груп за цією ознакою треба використовувати непараметричні методи.

### 1.3.2 Аналіз природних мов

Одним з найважливіших моментів у розвитку медичних інформаційних систем є те, що великі обсяги паперових документів треба було переводити, або по-іншому кажучи відцифровувати у електронні. Така задача з одного боку виглядає занадто складною для реалізації, але з іншого були спроби її втілити у життя, при чому більш менш вдало.

Зокрема, у всесвітній мережі Інтернет є відомі проекти Open Library та Google Books, що мають мільйони відсканованих повнотекстових книг, які додають користувачі цих сервісів. Для цього використовують обробку природніх мов. Існує декілька підходів до виконання таких задач, вони приведені нижче:

- лінгвістичний підхід;
- статистичний підхід;
- символьний підхід;
- коннективістський підхід;
- метод допоміжних векторів;
- прихована марковська модель;

- умовні випадкові поля;
- n-грамні моделі.

У сфері охорони здоров’я також розповсюджена проблема неструктурованих даних. Різноманітні зображення медичних звітів, дані електронної пошти з вільним текстом, голосові записи взаємодій представляють великі труднощі та перепони для традиційного інструментарію аналізу, але тут на допомогу приходять алгоритми обробки природних мов.

Використовуючи їх, алгоритми машинного навчання можуть зображення тексту переводити у документи, які можна редагувати; з’являється можливість витягати семантичне значення або обробляти пошукові запити, які написані звичайним текстом для того щоб повернути точні результати. Можна навести приклад та принцип роботи такої системи. Системи «питання-відповідь» дозволяють користувачам формувати запити на пошук у звичайній формі у вигляді питань на природній мові та отримувати відповіді на них. Принципи роботи таких систем полягає в наступному: на вхід подається сигнал та речення сформульоване на природній мові. Далі цей текст проходить автоматичну обробку, а саме: попередню обробку, витягування сутності, розбиття тексту на фрагменти, токенізація, семантичний та морфологічно-синтаксичний аналіз.

Перевага вищезазначених систем «питання-відповідь» у тому, що результатом відповіді на питання, задане на природній мові є не тільки набір

документів та абзаців, а саме точна відповідь на питання. Є можливість також переглянути й документи на базі яких сформована найточніша відповідь.

### 1.3.3 Прогнозування та підтримка прийняття рішень

Можливість знаходити та витягати певну інформацію з надвеликих обсягів даних має важливе значення для підтримки медичних рішень та створення прогнозів, де технології машинного навчання також сьогодні пророблюють собі шлях. Здібність оперативно виявляти та усувати ризики може значно покращити результати для пацієнтів з різною кількістю як поведінкових, так і клінічних станів.

Прикладом може слугувати Каліфорнійський університет та компанія General Electrics Healthcare, на базі яких розроблюється бібліотека алгоритмів прогнозної аналітики для пацієнтів з травмою, щоб прискорити надання швидкої медичної допомоги. Дослідники з університету Карнегі-Меллона та Медичної школи Уейл Корнелл також практикують використання машинного навчання задля розпізнавання варіацій моделей витрат та створення шляхів керування хронічними захворюваннями, знижуючи витрати для пацієнтів.

Фінансові та клінічні програми – це всього-на-всього перша стадія розвитку машинного навчання. Перш за все, конфіденційність та кібербезпека клієнтів медичних закладів є критично важливими для кожного постачальника медичних IT-послуг, а методи інтелектуального аналізу та алгоритми можуть бути найкращим способом зміцнити їх надійність.

Можливість майже стовідсотково зберігати ці дані з більшою чутливістю, може призвести до важливої перемоги в галузі.

### 1.4 Постановка задачі

Сьогоднішні технології не стоять на місці та поступово розвиваються, важко навіть уявити з яким прискоренням модернізуються способи та методи рішення тих чи інших задач. Звідси випливає, що потреба у навчанні нових та підтримці існуючих спеціалістів, завжди є відкрита. У мережі Інтернет є багато інформації про МІС та методологій по створенню цих систем. Досить важливо є й те, що велика кількість наборів даних для аналізу теж є у вільному доступі. Набір даних, або іншими словами дата сет, який використовується у цій роботі, створений для розробки та дослідження методів інтелектуального аналізу даних для діагностування ішемічної хвороби серця.

Тому для створення такої медичної інформаційної системи потрібно розв’язати такі задачі:

- розглянути існуючі базові методи, порівняти їх переваги та недоліки;
- дослідити метрики та зробити їх опис, їх актуальність та важливість для оцінки якості результатів;
- провести порівняльний аналіз підсумків та обрати найбільш придатні базові методи і моделі;

розробити власну модифікацію метода, який дозволить підвищити якість результатів оцінки, шляхом надання переваги найбільш придатній метриці.

## 1.5 Висновок до розділу

Сучасний світ неможливо уявити без медицини, вона була, є і буде одним з найбільших напрямків науки. Важливо зазначити, що успіх у дуже великій мірі залежить від того, як вчасно людина звернеться до лікарів і як точно буде поставлений їй діагноз. Лікарі теж люди, тому можуть допускати помилки через стрес, халатність, виробничу втому, тобто людський фактор ніхто не відміняв. У машин немає людського фактору, але якість результатів безпосередньо залежать від грамотності їх розробки та підтримки у робочому стані.

Серед проблем, які можуть виникнути у медичних інформаційних системах, можна підкреслити причину, яка криється у самих даних, якими ця ІС оперує. Правильна інтерпретація даних це запорука успіху номер один. Тому недивно, що сьогодні існує багато технологічних рішень цієї проблеми, які базуються на основі штучного інтелекту та машинного навчання. Зокрема, усім відомі корпорації Microsoft і Amazon вже зайняли цю нішу на ринку передових технологій.

# 2 МАТЕМАТИЧНІ МОДЕЛІ, МЕТОДИ ДЛЯ ВИРІШЕННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ КЛАСИФІКАЦІЇ

## 2.1 Попередній аналіз і обробка даних

Перша проблема з якою стикаються при обробці даних – це їх неправильний формат, дані представлені здебільшого «сирими» та використання таких наборів може призвести до неправильних обрахунків. Тому інтелектуальний аналіз даних включає у себе методи, які передбачають перетворення неготових даних у більш зрозумілу форму для подальшої роботи з ними. Дійсні дані, які містяться там часто трапляються неповними, суперечливими та містити багато пропущених значень.

Під час попередньої обробки даних проходять деякий ряд взаємодій, таких як: очищення даних, їх інтеграція, трансформація, скорочення та дискретизація. Велика кількість методів багатовимірного статистичного аналізу даних, зокрема факторний, регресійний, кластерний тощо, потребують відсутності порожніх значень при аналізі. Але як зазначено вище, в реальності набори даних бувають неповними дуже часто. Причин є дуже багато: можна вказати на те, що людина не ввела їх з різних міркувань або ж дані просто пошкодилися в процесі. Тому аналіз даних з відсутніми елементами покаже неповну картину ситуації, видасть помилкові висновки, а інколи, рішення прийняті на підставі таких висновків, можуть призвести до фатальних наслідків.

Дані, які можуть бути непридатними для обробки та аналізу, називають «брудними даними». Одні з найпоширеніших видів брудних даних є: відсутні дані, шуми і викиди, дублікати даних.

### 2.1.1 Відсутні дані

Трапляється, що деякі елементи даних можуть бути пропущені, причинами цього можуть слугувати:

- працівник, який вносив дані, допустив помилку;
- деякі ознаки не підходять для характеристики певних об’єктів;
- на той момент часу не було можливості для збору даних;
- дані взагалі не були зібрані.

Тому пропонується такі способи вирішення проблеми - наприклад ігнорування. Такий метод, звісно, є найелементарнішим для обробки порожніх елементів. При тому, якщо кількість відсутніх даних є надвеликою чи шаблон пов’язаний з неідентифікованим коренем формулювання проблеми – це метод використовувати не рекомендується взагалі.

Можна звернути увагу на метод заповнення пропущених елементів вручну. Таких спосіб є одним з найкращих, але при великих обсягах відсутніх значень, такий метод може зайняти велику кількість часу на виконання і ефективність його впаде.

Тоді звернемо увагу на метод обчислення значень – пропущенні дані можуть бути заповнені за допомогою обчислення середнього арифметичного

або медіани спостережуваних заданих значень. Треба пам’ятати, що такий метод може призвести до зміщення у вибірці, оскільки обрахункові значення не будуть точними щодо спостережуваних даних.

Залишається ще один спосіб – заміна порожніх елементів на деякі певні значення, які витікають з інших показників ознак об’єктів у наборі.

### 2.1.2 Шуми і викиди

Перш за все, треба дати означення поняттям шум і викид. Викид – це різко одмінні один від одного спостереження або об’єкти в певному наборі даних, тобто такі об’єкти, ознаки яких вибивають ці об’єкти із загальної картини. Вони є розповсюдженою проблемою при аналізі даних. Шуми й викиди можуть бути окремими спостереженнями або об’єднаннями в певні групи. Тому існує поширений метод проведення двоетапного аналізу даних – з викидами та з їх відсутністю та зрівнюванні отриманих результатів.

Одноразовий випадок неточного вимірювання може бути допустимим та пов’язаним з властивою технічною помилкою приладу яким вимірюють. Але деякі помилки заслуговують більш особливої уваги, та визначення які з цих помилок є більш важливими – сильно залежить від обраного досліду. Під час клінічних досліджень, неточності, які потребують негайного виправлення, включають у себе невірні дані щодо статі, дати народження чи помилки медичного аналізу тощо. Можна привести приклад: при аналізі який пов’язаний з дослідженням харчування, помилкова дата призводить до

невірного віку, а це призводить до помилкового обрахунку ваги за віком і до неправильного результату класифікації суб’єктів як нездорової ваги.

Проблема шуму та викидів у сфері інтелектуального аналізу даних зустрічається надто часто. Викиди це й окремі спостереження, так і об’єднання у цілі групи. Тому завдання спеціаліста-аналітика – їх виявити та оцінити степінь їх впливу на результати подальшого дослідження.

Щоб показати приклад виявлення викидів, часто вдаються до візуалізації. На рис. 2.1 зображено очевидний приклад викиду у вибірці даних.

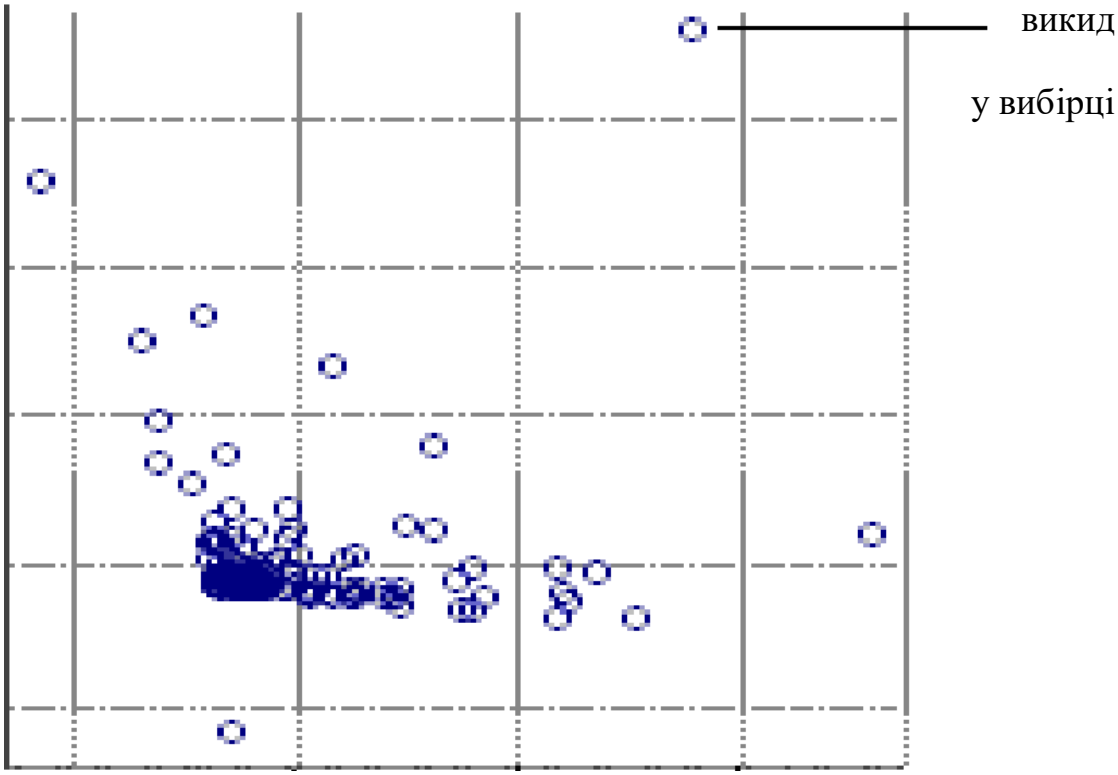


Рис. 2.1 – діаграма з зображенням викиду

Окрему проблему становить підводний камінь, під назвою «inlier». Це такі помилкові дані, які були внесені невірно, але при цьому вони розташовуються у межах очікуваного діапазону та не можуть бути ідентифіковані як викиди. Деяких таких замаскованих даних можна виявити досліджуючи історію кожної точки даних або повторним розрахунком, але такий аналіз нечасто можливий. Тому є змога змінити або оцінити набір об’єктів-помилок для оцінки частоти цих помилок.

Можна зробити підсумки, що висновки обраховані на підґрунті аналізу брудних даних не можуть мати достатньої довіри для прийняття рішення. Однак присутність цих брудних даних не завжди означає необхідність їх очищення. Поняття про доцільність вибору між наявність брудних даних та витратами на аналіз, повинне бути завжди.

### 2.1.3 Дублювання даних

Не секрет, що будь-яка вибірка може включати і дублікати даних. Дублікати – це об’єкти, усі властивості яких збігаються. Треба мати на увазі, що присутність дублікатів у дата сеті може бути способом підвищення значущості певних записів. Тобто. Така необхідність деколи виникає для особливого виділення деяких записів з вибірки. Але в основному, дубльовані об’єкти є результатом допущених помилок при створенні. Для усунення такого моменту існує два способи виправлення дублікатів:

1. Видалення усієї групи об’єктів, що дублюються. Цей метод використовується тільки тоді, коли присутність дублікатів спричиняє недовіру до наявної в них та знецінює її.
2. Заміна деякої групи дублікатів на один унікальний об’єкт.

### 2.1.4 Методи очистки даних

Процес знаходження та видалення невідповідностей та помилок у даних з метою їх покращення, називається очисткою даних.

Проблеми з якість цих даних можуть траплятися й в окремих вибірках даних, зокрема у файлах або базах даних. При інтеграції нескінченного числа джерел інформації, серед яких глобальні інформаційні Інтернет-системи або інтегровані системи баз даних, потреба в очищенні даних різко збільшується. Причиною слугує те, що джерела у більшості випадків мають розрізнені дані та виключення інформації, що дублюється. Перетворення створюється користувачем в інтерактивному режимі, або ж у формі бібліотеки правил.

Зараз існує багато методів очищення наборів даних, але важливим залишається те, що вони повинні задовольняти таким критеріям:

1. Метод повинен виявляти та видаляти усі невідповідності і помилки незважаючи чи це окремі джерела, чи це інтеграція декількох джерел даних.
2. Повинен постійно підтримуватися деяким інструментарієм, задля скорочення обсягів програмування і ручної перевірки.

3. Процес очистки не повинен відбуватися у відриві від зв’язаних зі схемою перетворення даних, виконуваних на базі складних метаданих.
4. Інфраструктуру технологічного процесу необхідно інтенсивно підтримувати для сховищ даних.

Також не буде зайвим перерахувати основні етапи очищення даних:

1. аналіз даних;
2. визначення порядку й правил перетворення даних;
3. підтвердження;
4. саме перетворення;
5. протитечія очищених даних;

Не дивно, що сьогодні очищення даних користується великим попитом. Великий ряд дослідницьких груп займається проблемами, пов’язаними з очищенням даних, включаючи специфічні підходи до Data Mining і перетворення даних на основі зіставлення схеми.

### 2.1.5 Методи заповнення пропущених елементів

Проблему пропущених даних може охарактеризувати така причина, як людський фактор, наприклад помилки в анкетах, неуважність респондентів тощо. У підсумку утворюється неповний масив даних з пропущеними елементами. Тому розроблено багато методик відновлення цих значень.

В основному використовують методи заповнення пропусків вже після етапу забору інформації: заповнення середніми значеннями, пропорційне розміщення спостережень з пропущеними, обрахунки за допомогою регресійних моделей тощо. Для прикладу, можна привести такі: метод Бартлетта, коли у випадку активного експерименту значення факторів задаються дослідником і пропуски в вихідних даних помічаються частіше, ніж у вхідних факторах; або за допомогою методу максимізації очікувань, який дозволяє не тільки відновлювати порожні значення з використанням двоетапного ітеративного алгоритму, а й оцінювати середні значення, кореляційні і коваріаційні матриці для кількісних змінних.

### 2.1.6 Методи нормування набору даних

Кожний набір вихідних даних, який завантажується в аналітичну програму, охарактеризовується деяким набором властивостей, які впливають на ефективність роботи моделі і знижують вірогідність отримання коректних результатів аналізу. Дані можуть бути представлені у форматі, з якими не працює певний алгоритм, можуть бути невпорядкованими. Для рішення цього моменту використовують трансформацію даних, задля приведення їх до деякого підходящого формату або виду.

Можна перерахувати такі методи трансформації:

- сортування;
- злиття;

- налаштування даних;
- нормалізація даних;
- квантування даних;
- дискретизація даних;
- додавання нових атрибутів, які обчислені на підставі вже наявних в даних;

## 2.2 Вибір базових класифікаторів

### 2.2.1 Постановка задачі класифікації

Для виконання аналітичних задач використовують навчання без вчителя, з вчителем, навчання з підкріпленням та напівавтоматичне. У цій роботі предмет вивчення торкається такого типу машинного навчання, як навчання з вчителем, а саме – бінарну класифікацію.

Задача навчання по прецендентам виглядає так:

Припустимо, що $X$ – множина об’єктів, $Y$ – множина відповідей, $y: X \to Y$ – залежність, яка невідома.

Дано:

$\{ x_1, \dots, x_l \} \subset X$ – навчальна вибірка;

$y_i = y(x_l), i = 1, \dots, l$ – відомі відповіді;

Знайти:

$a: X \to Y$ – алгоритм, функцію прийняття рішення, що $y$ приближує на всій множині $X$.

Тепер дослідимо, що таке об’єкти та відповіді.

$f_j: X \to D_j, j = 1, \dots, n$ – ознаки об’єктів.

Типи ознак:

- $D_j = \{0,1\}$ – бінарна ознака $f_j$;
- $|D_j| < \infty$ - номінальна ознака $f_j$;
- $|D_j| < \infty$, $D_j$ впорядкована – порядкова ознака $f_j$;
- $D_j = R$ – кількісна ознака $f_j$;

Вектор $(f_1(x), \dots, f_n(x))$ називають вектором ознак об’єкта $x$.

Також розглянемо матрицю «об’єктів ознак»

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix}$$

Відповіді можуть бути задані наступним способом:

Для задач класифікації:

- $Y = \{-1, +1\}$ – класифікація на 2 класи;
- $Y = \{1, \dots, M\}$ – класифікація на М класів, які не перетинаються;
- $Y = \{0,1\}^M$ – на М класів, що можуть перетинатися;

Для задач регресії:

- $Y = R$ або $Y = R^m$;

Для задач ранжування:

- $Y$ – скінченна впорядкована множина;

Метод навчання відображається так:

$$\mu: (X \times Y)^l \to A,$$

що, любому набору $X^l$ ставить у відповідність деякий алгоритм $a \in A$.

Задачі з машинного навчання по прецендентах складаються з двох етапів: це навчання та тестування.

Для того, щоб позначити функціонали якості, використовують поняття функції втрат $\mathcal{L}(a, x)$ - це величина помилки алгоритму $a \in A$ на об'єкті $x \in X$.

Як правило, для задач класифікації використовують $\mathcal{L}(a, x) = [a(x) \neq y(x)]$ – індикатор помилки. Але у випадку рішення задач регресії використовують або квадратичну похибку $(\mathcal{L}(a, x) = (a(x) - y(x))^2)$ або абсолютне значення помилки $(\mathcal{L}(a, x) = |a(x) - y(x)|)$.

Тому виходячи з цього, задача навчання перетікає у задачу оптимізації за допомогою введення поняття емпіричного ризику як функціоналу якості алгоритму $a$ на $X^l$.

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} \mathcal{L}(a, x_i),$$

$$\mu(X^l) = \underset{a \in A}{\arg\min}\, Q(a, X^l)$$

Можна зробити висновки, що класифікація – один із розділів напряму машинного навчання, який вирішує задачу, коли маємо множину об'єктів, які розділені на класи за деякою чи за декількома ознаками. Більш того, така скінчена множина об'єктів, де відомо яким саме класам вони належать. Таку множину об'єктів називають навчальною вибіркою. При цьому невідомо до якого класу відноситься решта інших об'єктів.

Далі необхідно побудувати алгоритм, який буде спроможний класифікувати любий з об'єктів вихідної множини. На виході буде видана відповідь алгоритму, що вказує номер чи назву класу, якому належить цей об'єкт – це буде вважатися класифікацією об'єкта.

В машинному навчанні, завдання класифікацій відносять до задач навчання з вчителем. Нижче приведені типи, на які їх поділяють:

- тип вхідних даних(предиктори): опис ознак, часові ряди, матриця відстаней, зображення тощо.
- тип класів: нечіткі класи, класи які перетинаються та не перетинаються, двокласова та багатокласова.

### 2.2.2 Лінійний класифікатор

У напрямку машинного навчання стоїть задача статистичної класифікації з застосуванням характеристик об’єкту для впізнавання класу або групи класів. Лінійний класифікатор виконує це завдання за допомогою ухвалення рішення про класифікацію об’єкта на основі значення лінійної комбінації характеристик. Вектор ознак – це характеристики цих об’єктів, які представляються у векторі, по-іншому їх також називають значенням ознак. Такі класифікатори, які розділяють простір ознак лінією, або площиною в просторі, розмірність якого більше двох, називають лінійними класифікаторами.

Нижче розглянуто декілька прикладів таких класифікаторів.

**Одношаровий персептрон**

У такому класифікаторі використовується апроксимація сигмоїдною функцією:

$$[M < 0] \leq \frac{2}{1 + e^{\alpha M}}$$

де $\alpha$ – параметр, який задається апріорі. Тому потрібно використовувати додаткові евристики для зменшення перенавчання та вибивання із локальних мінімумів.

**Логістична регресія**

В основі логістичної регресії лежить сигмоїдна функція, яка була розроблена аналітиками та статистами для дослідження росту популяції видів у природі. Така функція має вид латинської літери «S» та приймає значення від 0 до 1, не включаючи їх самих.

$$\sigma = \frac{1}{1 + e^{-z}}$$

Апроксимація порогової функції втрат відбувається неперервною оцінкою зверху таким способом, а навчання проводять методом градієнтного спуску другого порядку:

$$[M < 0] \leq \log_2(1 + e^{-M})$$

### 2.2.3 Дерева рішень

Дерева рішень (англ. Decision Tree, DT) – один з представників логічних алгоритмів класифікації, він базується на пошуку кон'юнктивних закономірностей. Дерево рішення – це кінцевий зв'язний граф з множиною вершин V, що не має циклів та який має виділену вершину $v_0 \in V$, в яку не входить жодне з ребер. Така вершина називається коренем дерева. А вершина яка не має ребер, що з неї виходять, називається листом або термінальною. Усі інші вершини називаються внутрішніми.

Під поняттям бінарне дерево мається на увазі те, що з будь-якої її внутрішньої вершини виходить строго два ребра, які зв'язують кожну внутрішню вершину $v$ з лівою дочірньою вершиною $L_v$ та з правою дочірньою вершиною $R_v$. У такому класифікаторі в кожній вершині прописаний предикат $\beta v: X \to \{0, 1\}$, а кожному листу відповідає мітка класу. Також застосовується таке поняття як ентропія яка визначається таким чином:

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i,$$

де $p_i$ – це ймовірність знаходження системи в стані $i$.

Важливо указати на те, що в основі поширених алгоритмів лежить принцип жадібної максимізації приросту інформації. Тобто на кожному кроці обирається та ознака, при якої приріст ентропії виявляється найбільшим, якщо її розбити. Після цього цю процедуру повторюють стільки разів, поки

кількість інформації не буде дорівнювати нулю або не буде менше заданого числа.

Також виділяють серед інших критеріїв якості розбиття в задачах класифікації:

- помилку класифікації;

$$E = \max_k p_k$$

- невизначеність Джині: $\mathrm{G} = 1 - \sum_k (p_k)^2$ . Мається на увазі максимізація числа пар об’єктів одного класу, які знаходяться в одному під-дереві.

Важливо також перерахувати переваги та недоліки використання дерев рішень. Переваги:

- потребують відносно невеликої попередньої обробки даних;
- на якість роботи не впливають нелінійні залежності між параметрами;
- можуть працювати з числовими та якісними або категоріальними ознаками;
- зрозумілий алгоритм, який легко можна візуалізувати.

Недоліки використання:

- при невеликих відхиленнях в даних, стають нестійкими;
- нестійкі до перенавчання;

### 2.2.4 Наївний баєсів класифікатор

Це ймовірнісний класифікатор, який використовує Теорему Баєса та дає можливість визначити ймовірність належності спостереження або елемента вибірки до одного з класів при наївному припущенні незалежності змінних. Тому, якщо на базі значень змінних можна точно визначити до якого з класів приналежить спостереження або елемент вибірки, то баєсів класифікатор покаже ймовірність належності до цього класу.

Якщо йдеться про проміжні випадки, коли спостереження може з різною ймовірністю належати до того чи іншого з класів, результат роботи класифікатора буде виданий в якості вектора, чиї компоненти є ймовірностями належності до певного класу.

Класифікація за допомогою цього метода проводиться в припущенні, що кожний з об’єктів $x \in X$ описується $n$ незалежними ознаками. Тобто, функція правдоподібності для класів розпадається на добуток. Сам класифікатор може бути параметричним, або непараметричним, це залежить від того, яким методом одновимірні щільності відновлюються.

Тепер важливо підкреслити переваги даного методу, серед них: малі обчислювальні витрати під час навчання та класифікації; проста реалізація. У рідких випадках, коли ознаки взаємонезалежні, наївний баєсів класифікатор – оптимальний.

Серед недоліків можна визначити відносно низьку точність класифікації в більшості реальних задач.

## 2.3 Метрики оцінки роботи класифікаторів

Після відбору ознак, проектування, побудови моделі й отримання результатів, що представлені у виді ймовірностей або класів, наступним етапом повинна бути перевірка ефективності моделі за допомогою спеціалізованих метрик та тестового набору даних. В залежності від поставленої задачі, спеціаліст обирає метрики які будуть використані в подальшій роботі.

Вибір метрик має важливий вплив на якість порівняння та оцінки алгоритмів між собою. Матриця помилок використовується для оцінки якості роботи класифікаторів, і для задачі бінарної класифікації має такий вид (рис. 2.2):

| | | | |
|---|---|---|---|
| Predicted values | True Values | y = 1 | y=0 |
| | a(x) = 1 | True Positive (TP) | False Positive (FP) |
| | a(x) = 0 | False Negative (FN) | True Negative (TN) |

Рис. 2.2 – Матриця помилок

Представимо, що розглядається задача медичного діагностування, де $y = 1$ означає що обстежуваний пацієнт хворий, а $y = 0$ – пацієнт здоровий.

True Positive (TP) – буде означати випадок, де реальне значення і відповідь алгоритму дорівнювали – теж 1.

True Negative (TN) – алгоритм у здорової людини не виявив хвороби.

False Positive (FP) – помилковий випадок, де реальне значення дорівнювало 0 – тобто відсутності хвороби, а прогнозоване – 1, її присутності. Помилка вважатиметься позитивною, адже людина здорова, а результат невірний, тому сама помилка несе менше загрози.

False Negative (FN) – також помилковий випадок, але тепер ціна помилки буде високою для медичної сфери, так як алгоритм видав справді хвору людину за здорову.

Не секрет, що хотілося би того, щоб наша інформаційна система видавала завжди вірний безпомилковий результат, але у реальному житті таке маловірогідне, тому кожна модель повинна мінімізувати кількість хибних результатів.

### 2.3.1 Точність (Accuracy)

Точність процесу класифікації рахується як співвідношення між вірними відповідями й усією кількість відповідей.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy – це ефективна метрика якості для роботи з збалансованими класами даних. Але гарні результати, які будуть показані є неправильними, якщо кількість об’єктів одного класу досить більше ніж іншого. Наприклад, якщо взяти вибірку зі 100 пацієнтами, серед яких буде 97 здорових, а хворих 3, то незважаючи на погану якість моделі, алгоритм всіх діагностує як здорових, при чому значення Accuracy буде 97%.

### 2.3.2 Чіткість (Precision)

Формула обрахунку виглядає наступним чином:

$$Precision = \frac{TP}{TP + FP}$$

Чіткість – метрика, яка показує на частку людей, у яких була діагностована хвороба вірно. Якщо знову привести приклад, де зі 100 пацієнтів хворими є тільки 3, то якщо тепер наша неякісна модель у кожному випадку діагностує хворобу, то тепер значення Precision дорівнюватиме 3%.

### 2.3.3 Повнота (Recall)

Повнотою називають міру, яка дає повну інформацію про те, який частка хворих пацієнтів була діагностована як хворі.

$$Recall = \frac{TP}{TP + FN}$$

Приведемо знову приклад з вибіркою, де зі 100 пацієнтів – 3 насправді хворі. Нехай алгоритм дає результат, що кожний випадок це наявність хвороби, тому у чисельнику та знаменнику буде значення 3, тому і повнота (Recall) моделі буде 100%, а чіткість (Precision) при цьому становитиме 3%.

Це ще раз доводить те, що ми завжди маємо бути уважними при виборі та застосуванні метрик якості, так як не у всіх випадках отриманий результат буде чітко оцінювати ефективність той чи іншої моделі.

### 2.3.4 Специфічність (Specificity)

Формула обрахунку:

$$Specificity = \frac{TN}{TN + FP}$$

Дана метрика указує на частку відповідей, які були класифіковані вірно з негативним результатом.

### 2.3.5 F-міра

Так як використання Precision та Recall не завжди є найкращим варіантом через їх похибність. Доречніше використовувати таку метрику:

$$F = \frac{2 * precision * recall}{precision + recall}$$

## 2.4 Висновок до розділу

У цьому розділі було розглянуті питання постановки завдання для машинного навчання та бінарної класифікації. Також був проведений огляд поширених моделей класифікації, зокрема логістична регресія, одношаровий персептрон, наївний баєсів класифікатор, дерева рішень; детально розглянуто їх недоліки та переваги. Також не менш важливо, що було звернено увагу на методи оцінки роботи якостей моделей, а саме точність, чіткість, повнота та інші.

# 3 ОЦІНКА ЯКОСТІ РОБОТИ ІНФОРМАЦІЙНОЇ СИСТЕМИ ТА БАЗОВИХ КЛАСИФІКАТОРІВ ДЛЯ ДІАГНОСТУВАННЯ ПАЦІЄНТА

## 3.1 Попередня обробка навчальних даних

### 3.1.1 Опис вибірки

У даній роботі був використаний розширена вибірка навчальних даних Z-Alizadeh Sani. Вона включає дані про 303 пацієнтів або об'єктів, які описуються 56 атрибутами. Ці атрибути поділені на 4 категорії: атрибути лабораторних досліджень і знімків, результати огляду та симптоми, демографічні, результати ЕКГ дослідження.

Нижче перераховані атрибути для кожного об'єкта:

- 'Age' – вік;
- 'Weight' – вага;
- 'Length' – зріст;
- 'Sex' – стать;
- 'BMI' – індекс маси тіла;
- 'DM' – хворий на сахарний діабет (так/ні);
- 'HTN' - має гіпертонію (так/ні);
- 'Current Smoker' – палить (так/ні);
- 'EX-Smoker' – чи палив в минулому (так/ні);
- 'FH' - сімейна гіперхолестеринемія (так/ні);
- 'Obesity' – страждає від ожиріння (так/ні);
- 'CRF' – хронічна ниркова недостатність (так/ні);

- 'CVA' – цереброваскулярна аварія (так/ні);
- 'Airway disease' – хвороба дихальних шляхів (так/ні);
- 'Thyroid Disease' – захворювання щитовидної залози (так/ні);
- 'CHF' – застійна серцева недостатність (так/ні);
- 'DLP' – дисліпопротеїнемія (так/ні);
- 'BP' – кров’яний тиск;
- 'PR' – протромбіновий коефіцієнт;
- 'Edema' – наявність набряків (так/ні);
- 'Weak Peripheral Pulse' – слабкий периферійний пульс (так/ні);
- 'Lung rales' – хрипи в легенях (так/ні);
- 'Systolic Murmur' – систолічний шум (так/ні);
- 'Diastolic Murmur' – діастолічний шум (так/ні);
- 'Typical Chest Pain' – типовий біль у грудях (так/ні);
- 'Dyspnea' – задишка (так/ні);
- 'Function Class' – функціональний клас (4 типи);
- 'Atypical' – нетиповий об’єкт (так/ні);
- 'Exertional CP' – фізичний церебральний параліч (так/ні);
- 'Q Wave' – хвилі типу Q (характеристика ЕКГ);
- 'St Elevation' – сильний підйом;
- 'St Depression' – сильна депресія;
- 'Tinversion' – характеристика ЕКГ;
- 'LVH' – гіпертрофія лівого шлуночка (так/ні);

- 'BBB' – гематоенцефалічний бар'єр (немає/лівий/правий);
- 'FBS' – рівень цукру у крові;
- 'CR' – умовний рефлекс;
- 'TG' – тигроглобулін;
- 'LDL' – ліпопротеїн низької щільності;
- 'HDL' – ліпопротеїн високої щільності;
- 'BUN' – азот сечовини крові;
- 'ESR' – електронний спіновий резонанс;
- 'HB' – гемоглобін;
- 'K' – калій;
- 'Na' – натрій;
- 'WBC' – білі кров'яні клітини;
- 'Lymph' – лімфа;
- 'Neut' – бікарбонат натрію;
- 'PLT' – пробний тест на лімфоцити;
- 'EF-TTE' – ехокардіограма;
- 'Region RWMA' – Регіон аномального руху стінок (ехокардіограма): 0-5;
- 'VHD' – клапанна хвороба серця (відсутня або одна з трьох степеней серйозності захворювання);
- 'Cath' – індикатор наявності ішемічної хвороби серця, цільова змінна.

### 3.1.2 Використання попередньої обробки даних

Перед тим, як перейти до аналізу вибірки даних, ці самі дані потрібно попередньо обробити. А саме, перетворити категоріальні дані в числові, їх трансформувати та відібрати ознаки. Також такий метод має назву feature selection.

Оскільки присутність нерелевантних ознак у вибірці може спричинити зменшення точності багатьох моделей, наприклад логістичну або лінійну регресію. Цей метод має декілька переваг, тому використання відбору ознак є обґрунтованим. Він підвищує точність аналізу, шляхом відсіювання даних, які можуть спричинити значну похибку. Також зменшує час налаштування моделі та знижує її перенавчання.

Метод feature selection розділяється на дві групи на підставі критеріїв оцінки: підходи, які використовують фільтрацію та ті, які використовують «обгортку». Наприклад способи фільтрації не залежать від методу навчання класифікаторів та застосовують методи вимірювання, такі як відстань, кореляція та послідовність у даних задля знаходження найбільш вдалої підмножини з цілого набору атрибутів. Але не дивлячись на швидкість роботи цього підходу, він зневажає кореляцією між продуктивність алгоритму індукції та обраною підмножиною. А в підході «обгортки» попередньо визначений алгоритм навчання застосовується для виміру ефективності узагальнення методом стохастичного пошуку для максимізації даних.

Важливо також звернути увагу на недоліки підходу, що використовує методи «обгортки». Одним з найбільших є вимоги для обчислення, так як усі можливі піднабори ознак мають бути оцінені з використанням класифікатора.

Причиною появи гібридного методу, у якому застосовується евристики для більш якісного процесу відбору релевантних ознак, стала недостатня надійність підходів фільтрації та висока обчислювальна вартість методів обгортки.

Нижче перераховані відомі методи відбору ознак:

- Відбір, який основується на аналізі однієї змінної та сили її зв’язку з цільовою;
- «Жадібні» методи. Це відбір, де дозволяється автоматизувати прийняття рішення щодо остаточної кількості змінних. Можна привести приклад методу Rfecv, який включає рекурсивне видалення атрибутів шляхом ранжування та використання крос-валідації.

Для того, щоб зменшити розмірність простору ознак застосовують лінійно-дискримінантний аналіз(LDA) і метод головних компонент(PCA). Існують алгоритми класифікації, наприклад дерев рішень, які роблять ранжування ознак за степенем їх важливості.

Такі дані теж можуть бути розглянуті як вхідні дані для лінійних методів класифікації.

Нижче приведена таблиця 3.1 відбору ознак.

| Feature selection approach | Number of features |
|---|---|
| Rfecv (logistic regression) | 33 |
| Decision trees | 13 |
| PCA | 20 |
| LinearSVC | 25 |

Таблиця 3.1 Результати feature selection

У даній роботі використовується два способи обробки даних: мінімаксне масштабування та стандартизація.

Мінімаксне масштабування трансформує кожну ознаку так, щоб її значення розташовувались у заданому діапазоні набору, тобто від 0 до 1. Також беручи до уваги те, що класи у вибірці незбалансовані, тому для уникнення втрат важливих даних, були обчислені вігові коєфіцієнти для кожного з класів.

## 3.2 Результати використання простих класифікаторів

Усі моделі класифікаторів потребували підбор параметрів та налаштування. Тому були використані крос-валідація на чотирьох блоках.

### 3.2.1 Результат використання логістичної регресії

Нижче у таблиці 3.2 наведені результати діагностування пацієнта або класифікації шляхом використання логістичної регресії.

| | **accuracy** | **precision** | **recall** | **specificity** | **f1 score** | **roc auc score** |
|---|---|---|---|---|---|---|
| | 0.87 | 0.89 | 0.93 | 0.72 | 0.91 | 0.83 |
| rescaled (minmax) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| rescaled (stand) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (decision trees) | 0.89 | 0.95 | 0.88 | 0.89 | 0.92 | 0.89 |
| feature selection (rfecv) | 0.85 | 0.9 | 0.88 | 0.78 | 0.89 | 0.83 |
| feature selection (f classif) | 0.84 | 0.87 | 0.91 | 0.89 | 0.79 | 0.67 |
| feature selection (mutual information) | 0.89 | 0.91 | 0.93 | 0.78 | 0.92 | 0.85 |
| feature selection (svm) | 0.87 | 0.89 | 0.93 | 0.72 | 0.91 | 0.83 |
| feature selection (pca) | 0.89 | 0.91 | 0.93 | 0.78 | 0.92 | 0.85 |
| feature selection (pca) | 0.89 | 0.93 | 0.91 | 0.83 | 0.92 | 0.87 |
| feature selection (decision trees) + rescaled (stand) | 0.9 | 0.97 | 0.88 | 0.94 | 0.93 | 0.91 |
| feature selection (rfecv) + rescaled (minmax) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (rfecv) + rescaled (stand) | 0.89 | 0.97 | 0.86 | 0.94 | 0.91 | 0.9 |
| feature selection (f classif) + rescaled (minmax) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (f classif) + rescaled (stand) | 0.98 | 0.93 | 0.91 | 0.83 | 0.92 | 0.87 |

Кінець таблиці 3.2

| | | | | | | |
|---|---|---|---|---|---|---|
| feature selection (mutual info) + rescaled (minmax) | 0.85 | 0.89 | 0.91 | 0.72 | 0.9 | 0.81 |
| feature selection (mutual info) + rescaled (stand) | 0.84 | 0.92 | 0.84 | 0.83 | 0.88 | 0.84 |
| feature selection (svm) + rescaled (minmax) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (svm) + rescaled (stand) | 0.85 | 0.9 | 0.88 | 0.78 | 0.89 | 0.83 |
| feature selection (pca) + rescaled (minmax) | 0.9 | 0.95 | 0.91 | 0.89 | 0.93 | 0.9 |
| feature selection (pca) + rescaled (stand) | 0.85 | 0.93 | 0.86 | 0.83 | 0.89 | 0.85 |

### 3.2.2 Використання баєсова наївного класифікатора

Нижче у таблиці 3.3 приведені результати роботи класифікатора за допомогою моделі NB.

Таблиця 3.3 – Результати роботи баєсова наївного класифікатора

| | **accuracy** | **precision** | **recall** | **specificity** | **f1 score** | **roc auc score** |
|---|---|---|---|---|---|---|
| | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| rescaled (minmax) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |

Кінець таблиці 3.3

| | | | | | | |
|---|---|---|---|---|---|---|
| feature selection (rfecv/ fclassif/ mutual-info) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (svm) | 0.84 | 0.84 | 0.95 | 0.56 | 0.89 | 0.75 |
| feature selection (decision trees) + rescaled (minmax) | 0.85 | 0.9 | 0.88 | 0.78 | 0.89 | 0.83 |
| feature selection (decision trees) + rescaled (stand) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (rfecv) + rescaled (minmax) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (fclassif) + rescaled (stand) | 0.89 | 0.91 | 0.93 | 0.78 | 0.92 | 0.85 |
| feature selection (mutual info) + rescaled (minmax) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (mutual info) + rescaled (stand) | 0.85 | 0.87 | 0.93 | 0.67 | 0.9 | 0.8 |
| feature selection (svm) + rescaled (minmax) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (svm) + rescaled (stand) | 0.87 | 0.89 | 0.93 | 0.72 | 0.91 | 0.83 |
| feature selection (pca) + rescaled (stand) | 0.84 | 0.85 | 0.93 | 0.61 | 0.89 | 0.77 |

### 3.2.3 Результати використання дерев рішень

Нижче приведені результати роботи класифікатора за допомогою дерев рішень у таблиці 3.4.

Таблиця 3.4 – Результат використання дерев рішень.

| | accuracy | precision | recall | specificity | f1 score | roc auc score |
|---|---|---|---|---|---|---|
| | 0.82 | 0.88 | 0.86 | 0.72 | 0.87 | 0.79 |
| feature selection (decision trees) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (rfecv) | 0.82 | 0.83 | 0.93 | 0.56 | 0.88 | 0.74 |
| feature selection (f-classif) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (mutual information) | 0.87 | 0.87 | 0.95 | 0.67 | 0.91 | 0.81 |
| feature selection (svm) | 0.84 | 0.87 | 0.91 | 0.67 | 0.89 | 0.79 |
| feature selection (decision trees) + rescaled (minmax) | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.84 |
| feature selection (decision trees) + rescaled (stand) | 0.85 | 0.9 | 0.88 | 0.78 | 0.89 | 0.83 |
| feature selection (rfecv) + rescaled (minmax) | 0.82 | 0.86 | 0.88 | 0.67 | 0.87 | 0.78 |

Кінець таблиці 3.4

| | | | | | | |
|---|---|---|---|---|---|---|
| feature selection (rfecv) + rescaled (stand) | 0.8 | 0.84 | 0.88 | 0.61 | 0.86 | 0.75 |
| feature selection (mutual info) + rescaled (minmax) | 0.72 | 0.84 | 0.74 | 0.67 | 0.79 | 0.71 |
| feature selection (mutual info) + rescaled (stand) | 0.85 | 0.87 | 0.93 | 0.67 | 0.9 | 0.8 |
| feature selection (svm) + rescaled (minmax) | 0.84 | 0.85 | 0.93 | 0.61 | 0.89 | 0.77 |
| feature selection (svm) + rescaled (stand) | 0.84 | 0.85 | 0.93 | 0.61 | 0.89 | 0.77 |
| feature selection (pca) + rescaled (minmax) | 0.7 | 0.8 | 0.77 | 0.56 | 0.79 | 0.66 |
| feature selection (pca) + rescaled (stand) | 0.69 | 0.8 | 0.74 | 0.56 | 0.77 | 0.65 |

### 3.3 Аналіз результатів роботи класифікаторів

При проведенні порівняльного аналізу моделей класифікаторів між собою, виходячи з вище приведених таблиць, має сенс розглядати оцінку повноти (recall store), тому що сам оцінка повноти обраховує частку правильно класифікованих об’єктів першого класу.

Розглядаючи природу обраного практичного завдання та відповіді класифікаторів у вигляді «нуля» та «одиниці», можна зробити висновок, що правильна класифікація об’єктів першого класу є найбільш важливою, так як хибний аналіз у хворої людини має більш негативний ефект та є небажаним.

Для наочного порівняння використаних моделей, нижче наведено найкращі результати у таблиці 3.5.

Таблиця 3.5 – Найкращі результати алгоритмів класифікації.

| **Метод** | **accuracy** | **precision** | **recall** | **specificity** | **f1 score** | **roc auc score** |
|---|---|---|---|---|---|---|
| Логістична регресія | 0.89 | 0.91 | 0.93 | 0.78 | 0.92 | 0.85 |
| Дерева рішень | 0.87 | 0.91 | 0.91 | 0.78 | 0.91 | 0.86 |
| Наївний байєсів класифікатор | 0.87 | 0.89 | 0.93 | 0.72 | 0.91 | 0.83 |

Дивлячись на таблицю, можна зробити висновки, що найбільшу recall score (оцінку повноти) має логістична регресія та наївний баєсів класифікатор. А найвищу чіткість (precision) має логістична регресія та дерева рішень. Показники f-міри найбільший також у логістичної регресії,

але не дуже відрізняється від інших моделей. Тому є сенс вважати логістичну регресією найкращим методом класифікації. Роздивляючись площу під roc-кривою, можна знайти найбільшу при використанні дерев рішень.

## 3.4 Висновок до розділу

У цьому розділі роботи було проведено значну кількість тестувань трьох типів класифікаторів, а також їх перевірка на якість під час вирішення задачі класифікації, а саме діагностування ішемічної хвороби серця. Для аналізу була використана відома вибірка даних Z-Alizadeh Sani. У ній містяться інформація про 303 пацієнтів, які описуються 56 атрибутами. За результатами роботи моделей був виконаний порівняльний аналіз, який показав недоліки та переваги кожного з методів. Було підмічено, що логістична регресія має найкращі показники.

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

Чорноморський національний університет імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

**Спеціальний розділ**

# ОХОРОНА ПРАЦІ

**до кваліфікаційної роботи**

на тему:

# ОХОРОНА ПРАЦІ З ПРАВИЛ ДОТРИМУВАННЯ ПОКАЗНИКІВ МІКРОКЛІМАТУ У РОБОЧИХ ПРИМІЩЕННЯХ

Спеціальність 122 «Комп'ютерні науки та інформаційні технології»

**122 – ДР – 402.21810206**

**Студент**_____С.Ю. Воздіцький

« 21 » червня 2022 р.

**Консультант** _____А.О. Алексєєва
ст. викладач
« 21 » червня 2022 р.

**Миколаїв – 2022**

# ВСТУП

Суттєвий плив на стан здоров’я працівника підприємства, його працездатність здебільшого впливає мікроклімат або метеорологічні умови у приміщеннях, під яким розуміють умови внутрішнього середовища цих приміщень, що впливають на тепловий обмін працюючих з оточенням. Ці умови визначаються поєднанням відносної вологості, швидкості руху повітря та температури поверхонь, що оточують людину, а також інтенсивністю теплового або інфрачервоного випромінювання.

Нормальне теплове самопочуття людини під час виконання будь-якої роботи може бути досягнуто за певної комбінації цих параметрів. Значення цих параметрів, які забезпечують найкраще самопочуття і найвищу працездатність людини, вважають оптимальними нормами мікроклімату.

Відхилення зазначених параметрів повітряного середовища від оптимальних норм створює несприятливі метеорологічні умови, що призводить до погіршення самопочуття, передвчасної втоми людини і зниження її працездатності.

**Метою роботи** є аналіз параметрів мікроклімату у офісному приміщені виробничого підприємства та надання рекомендацій для їх нормалізації.

Відповідно до мети виділені наступні етапи**:**

1. Створити опис офісного приміщення виробничого підприємства.
2. Визначити наявні параметри мікроклімату в офісному приміщенні.

3. Сформулювати рекомендації щодо нормалізації параметрів мікроклімату на робочому місті та офісному приміщені в цілому.

# 4 ОХОРОНА ПРАЦІ

## 4.1 Опис робочих приміщень

Офісне приміщення знаходиться на першому поверху, всередині виробничого приміщення. Офіс складається з чотирьох суміжних кімнат між якими немає дверей. Загальна площа приміщення – 127,6 кв. м. У приміщенні є шість вихідних дверей вироблених з металопластику розміром 800*2000. Також у приміщенні знаходяться шість металопластикових вікон, п’ять з котрих виходять у виробничі приміщення, а одне – на вулицю.

Загальний план офісного приміщення викладено на рис. 4.1

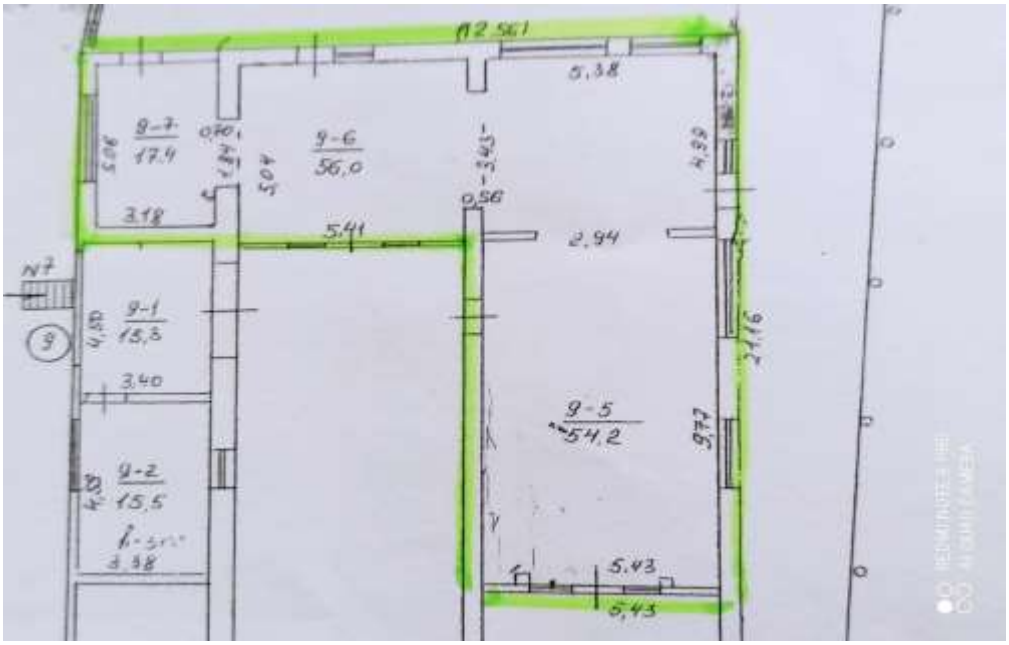


Рис.4.1. План приміщення підприємства ТОВ «МТРЗ»

Приміщення має сучасний офісний вигляд. Стіни пофарбовані у приємний колір – жовтий, покриття стін – декоративна штукатурка, стеля виконана у вигляді підвісної конструкції із синтетичного матеріалу білого

кольору з вставкою діодних світильників загальної кількістю 26 штук по 4 лампи кожний. Підлога виконана з ламінату світло-коричневого кольору.

Офісне приміщення розділене на 4 окремі зони, у яких розташовані різні відділи, а саме: 1) бухгалтерія; 2) комерційній відділ, 3) відділ договорів, 4) виробничий відділ та робоче місто системного адміністратора.

Взагалі в офісному приміщенні розташовано 18 робочих місць. Всі робочі місця обладнані сучасними комп'ютерами, принтерами. Для забезпечення нормального мікроклімату в офісі, приміщення у різних кінцях обладнані двома кондиціонерами, які в холодний період працюють на обігрів, а в теплий період на охолодження. Під стелею розташовано 4 телевізора, які постійно транслюють відображення камер відеоспостереження території всього виробничого підприємства.

Рис.4.2. Фото приміщення підприємства ТОВ «МТРЗ»

Так як приміщення офісне, то тут працюють робітники, які весь свій робочий час майже не покидають своїх робочих місць, тобто, категорія робіт працівників у даному приміщенні – це легкі фізичні роботи (Категорія I).

Офісні працівники виконують роботи сидячи і не потребують фізичного напруження, такі роботи відносяться до Категорії Іа, при якій витрата енергії працівника дорівнює 105-140 Вт (90-120 ккал/год).

Таблиця 4.1 Перелік обладнання в офісному приміщенні

| № п/п | Найменування | Од. вим. | Кількість |
|---|---|---|---|
| 1. | Стіл | шт | 18 |
| 2. | Крісло офісне | шт | 25 |
| 3. | Комп’ютер | шт | 18 |
| 4. | Принтер | шт | 10 |
| 5. | Кондиціонер | шт | 2 |
| 6. | Телевізор | шт | 4 |
| 7. | Шкап офісний | шт | 8 |
| 8. | Полка офісна | шт | 5 |

Усі працівники офісу знаходяться на своєму робочому місці понад 50% робочого часу, тому робочі місця в офісному приміщенні – є постійними.

Визначимо параметри мікроклімату в офісному приміщенні.

Для визначення, чи задовольняють фактичні параметри мікроклімату нормативним, нижче надамо оптимальні та допустимі параметри мікроклімату у виробничих приміщеннях:

Оптимальні величини температури, відносної вологості та швидкості руху повітря в робочій зоні виробничих приміщень приведені у таблиці 4.2.

Таблиця 4.2 Показники мікроклімату у приміщеннях

| Період року | Категорія робіт | Температура повітря | Відносна вологість | Швидкість руху, м/сек. |
|---|---|---|---|---|
| Холодний період року | Легка I а | 22-24 | 60-40 | 0,1 |
| | Легка I б | 21-23 | 60-40 | 0,1 |
| | Середньої важкості IIа | 19-21 | 60-40 | 0,2 |
| | Середньої важкості IIа | 17-19 | 60-40 | 0,2 |
| | Важка III | 16-18 | 60-40 | 0,3 |
| Теплий період року | Легка I а | 23-25 | 60-40 | 0,1 |
| | Легка I б | 22-24 | 60-40 | 0,2 |
| | Середньої важкості IIа | 21-23 | 60-40 | 0,3 |
| | Середньої важкості IIа | 20-22 | 60-40 | 0,3 |
| | Важка III | 18-20 | 60-40 | 0,4 |

Можна зробити висновок, що у різні пори роки є різні показники температурного режиму. Важливо дотримуватися цих цифр, так як їх норми прописані у державних санітарних нормах. Допустимі величини температури,

відносної вологості та швидкості руху повітря в робочій зоні виробничих приміщень приведені у таблиці 4.3.

Таблиця 4.3 Показники температури

| Період року | Категорія робіт | Температура, ˚С | | | | Відносна Вологість (%) на робочих місцях – постійних і непостійних |
|---|---|---|---|---|---|---|
| | | Верхня межа | | Нижня межа | | |
| | | На постійних робочих місцях | На непостійних робочих місцях | На постійних робочих місцях | На непостійних робочих місцях | |
| Холодний період року | Легка Іа | 26 | 26 | 21 | 18 | 75 |
| | Легка Іб | 24 | 25 | 20 | 17 | 75 |
| | Середньої важкості IIа | 23 | 24 | 17 | 15 | 75 |
| | Середньої важкості Іа | 21 | 23 | 15 | 13 | 75 |
| | Важка III | 19 | 20 | 13 | 12 | 75 |

Кінець таблиці 4.3

| | | | | | | |
|---|---|---|---|---|---|---|
| Теплий період року | Легка I а | 28 | 30 | 22 | 20 | 55-при 28 ˚С |
| | Легка I б | 28 | 30 | 21 | 19 | 60-при 27 ˚С |
| | Середньої важкості IIа | 27 | 29 | 18 | 17 | 65- при 26 ˚С |
| | Середньої важкості IIа | 27 | 29 | 15 | 15 | 70-при 25 ˚С |
| | Важка III | 26 | 28 | 15 | 13 | 75-при 24 ˚С |

У офісному приміщені виконано заміри параметрів мікроклімату, які наведено у таблиці 4.4.

Таблиця 4.4 Заміри показників мікроклімату

| | |
|---|---|
| Дата вимірювання | 27.05.2022 рік |
| Характеристика робочого міста | Постійне |

Кінець таблиці 4.4

| | |
|---|---|
| Енерговитрати організму | 120 ккал/год |
| Категорія робіт | Легка 1а |
| Період року | Теплий період |

## 4.2 Вимірювання показників робочих місць

Заміри параметрів мікроклімату виконувалися наступним обладнанням

1. Температуру повітря вимірювали електронним термометром;


Рис.4.3. Електронний термометр TOSOT для вимірювання температури повітря

Електронний термометр TOSOT використовується для вимірювання температури повітря в приміщенні. Важливо враховувати й похибки, які можуть бути при використанні такого типу термометрів. Наприклад, точності

ртутного термометра такий термометр не поступається, й похибка вимірювання двома приладами однакова й становить приблизно 0,1°С. Для того, щоб бути впевненим у правильності вимірювань, потрібно проводити перевірку термометрів не менше одного разу на рік.

2. Швидкість руху повітря вимірювали чашковим анемометром;

Рис.4.4. Чашковий анемометр Atmos для вимірювання швидкості руху повітря

Чашковий анемометр “Atmos” призначений для визначення швидкості вітру, температури та коефіцієнта охолодження вітром

3. Відносну вологість вимірювали гігрометром психометричним

Рис.4.5. Гігрометр психометричний для вимірювання відносної вологості

Гігрометр психрометричний призначений для виміру відносної вологості і температури повітря у приміщенні.

Таблиця 4.5. Показники вимір мікроклімату у приміщеннях

| Параметр мікроклімату | | | Задовольняє/ не задовольняє | Висновки | |
|---|---|---|---|---|---|
| Найменування | | Значення | | | |
| t, ˚C | фактична | 21 | Не задовольняє | Збільшити значення на | 1-4 ˚C |
| | | | | Зменшити значення на | |
| | оптимальна | 23-25 | | | |
| | допустима | 22-28 | | | |

Кінець таблиці 4.5

| | | | | | |
|---|---|---|---|---|---|
| W, % | фактична | 50 | Задовольняє | Збільшити значення на | - |
| | | | | Зменшити значення на | - |
| | оптимальна | 60-40 | | | |
| | допустима | 55 | | | |
| V, м/с | фактична | 0,3 | | Збільшити значення на | |
| | | | | Зменшити значення на | 0,1-0,2 |
| | оптимальна | 0,1 | | | |
| | допустима | 0,2-0,1 | | | |

Тобто, в офісному приміщені параметри мікроклімату не задовольняють вимогам ДСН 3.3.6.042-99. Для досягнення допустимих значень параметрів мікроклімату необхідно збільшити температуру повітря на 1-4 ˚С та зменшити швидкість руху повітря на 0,1-0,2 м/сек. Це можливо зробити піднявши на кондиціонерах бажану температуру повітря та зменшити швидкість роботи вентилятора на кондиціонері за допомогою пульта керування. Для рівномірного розподілення охолодженого чи теплого повітря по площі офісного приміщення, яке надають кондиціонери, на

підприємстві зроблено розсіювач повітря, це також вплине на швидкість руху повітря.

## 4.3 Висновки до розділу

В цьому спеціальному розділі з охорони праці було розглянуто важливість параметрів мікроклімату на робочому місці працівника офісу у виробничому приміщенні. Визначені категорії видів діяльності та характеристика робочих місць, від якої також залежать допустимі величини параметрів мікроклімату робочого місця в виробничому приміщенні. Після замірів фактичних величин параметрів мікроклімату виявлено, що в даному виробничому приміщенні температура повітря для працівників, які займаються легкою роботою Категорії Іа, та працюють на постійних робочих місцях є нижче норми, тобто її необхідно збільшити, а швидкість руху повітря навпаки більше норми, її необхідно зменшити.

Так як мікрокліматичні умови в виробничому приміщенні забезпечуються двома кондиціонерами, то виправити ситуацію можливо дуже швидко, потрібно підняти показник температури та контролювати його значення, та зменшити швидкість руху повітря, за допомогою зниження швидкості роботи вентилятора кондиціонеру. Це дозволить встановити в виробничому приміщенні оптимальні мікрокліматичні умови, які відповідають нормативним.

# ВИСНОВКИ

Під час виконання бакалаврської кваліфікаційної роботи було розглянуто що таке інформаційна система, основні етапи розвитку ІС, їх види та призначення. Згідно з темою даної роботи, було розглянуто питання медичних інформаційних систем та розроблено програму для аналізу даних для діагностування ішемічної хвороби серця у пацієнта. Також було вивчені основні класифікатори, їх переваги та недоліки. Усіх вищезазначених моментів було досягнуто завдяки:

- Аналізу актуальності, предмету та мети дослідження;
- На основі існуючих МІС та класифікаторів побудовано програму для реалізації інтелектуального аналізу даних та класифікації;
- Проведено попередню обробку та підготовку даних перед аналізом;
- Проведено навчання МІС;
- Вивчено результати аналізу та обрано найкращі моделі;

Для програмної реалізації медичної інформаційної системи було обрано мову програмування Python та використано існуючі бібліотеки для машинного навчання. Завдяки цьому було проведено навчання та аналіз даних на основі дата сету Z-Alizadeh Sani з відкритих джерел, який містить дані про пацієнтів та їх атрибути.

Бакалаврська кваліфікаційна робота складається з вступу, трьох розділів, висновку, переліку джерел, одного додатку та спеціальної частини

з охорони праці. Основна частина викладена на 55 сторінках тексту (без додатків), містить 9 рисунків, 11 таблиць та 23 джерел посилання.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Daniel M. Study Suggests Medical Errors Now Third Leading Cause of Death in the U.S. [Електронний ресурс] / Daniel M. – URL: https://www.hopkinsmedicine.org/news/media/releases/study_suggests_medical_errors_now_third_leading_cause_of_death_in_the_us

2. Фармацевтична Енциклопедія [Електронний ресурс]: / уклад. Стречень С.Б – URL: http://www.pharmencyclopedia.com.ua/article/2526/diagnostika

3. Gemp I. Automated Data Cleansing through Meta-learning [Електронний ресурс] / Gemp I., Theocharous G., Ghavamzadeh M., – Association for the Advancement of Artificial Intelligence, 2017– URL: https://people.cs.umass.edu/~imgemp/pubs/iaai_2017.pdf

4. Koktysh L. The state of the art in health data analytics [Електронний ресурс] / Koktysh L. – URL: https://www.scnsoft.com/blog/health-data-analytics-overview

5. Armstrong S. Data, data everywhere: the challenges of personalised medicine [Електронний ресурс] / Armstrong S. – URL: https://www.bmj.com/content/359/bmj.j4546

6. Bresnick J. IBM Watson Expands Role in Imaging Analytics, Population Health [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/ibm-watson-expands-role-in-imaginganalytics-population-health

7. Bresnick J. Imaging Analytics Get Big Data Boost from New Partnerships [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/imaging-analytics-get-big-data-boostfrom-new-partnerships

8. Bresnick J. Imaging Analytics Get Big Data Boost from New Partnerships [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/imaging-analytics-get-big-data-boostfrom-new-partnerships

9. Bresnick J. Data Governance Key to Hospital’s Natural Language Query Project [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/data-governance-key-to-hospitalsnatural-language-query-project

10. Bresnick J. Machine Learning, NLP Help with Physician Skill Benchmarking [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/machine-learning-nlp-help-withphysician-skill-benchmarking

11. Bresnick J. Mount Sinai Uses Machine Learning for Heart Imaging Analytics [Електронний ресурс] / Bresnick J. – URL: https://healthitanalytics.com/news/mount-sinai-uses-machine-learning-forheart-imaging-analytics

12. IBM Watson Health [Електронний ресурс] – URL: https://www.ibm.com/watson/health/oncology-and-genomics/oncology/

13. Predicting Response to Depression Treatment (PReDicT) project [Електронний ресурс] – URL: http://p1vital.com/ehealth/?p=376

14. Rajpurkar P. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks [Електронний ресурс] / Rajpurkar P., Hannun A., Haghpanahi M., Bourn C., NG A.Y. – Stanford – URL: https://arxiv.org/pdf/1707.01836.pdf

15. Mukherjee S. A.I. VERSUS M.D. What happens when diagnosis is automated? [Електронний ресурс] / Mukherjee S. – URL: https://www.newyorker.com/magazine/2017/04/03/ai-versus-md

16. Naumov L. Main Problems of Modern Medicine in Diagnostics and learning:
Ways of Optimal Solution [Електронний ресурс] / Naumov L.: Ana Kar Der. – 2001.– pp. 166-178 – URL: https://www.journalagent.com/anatoljcardiol/pdfs/AnatolJCardiol_1_3_166_178.pdf

17. Арустамов А. Статья: Предобработка и очистка данных перед загрузкой в хранилище [Електронний ресурс] / Арустамов А. — Режим доступа:http://sysdba.org.ua/proektirovanie-bd/etl/predobrabotka-iochistka-dannyih-pered-zagruzkoy-v-hranilische.html

18. Фоурино Р. Электронное качество данных: скрытая перспектива очистки данных [Електронний ресурс] / Фоурино Р.— Режим доступа: http://www.iso.ru/print/rus/document5820.phtml

19. Воронцов К.В. Машинное обучение: курс лекций [Електронний ресурс]
/ Воронцов К.В. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29

20. Воронцов К.В. Лекции по логическим алгоритмам классификации [Електронний ресурс] / Воронцов К.В. — Режим доступа: http://www.ccas.ru/voron/download/LogicAlgs.pdf

21. Дьяконов А. Введение в анализ данных и машинное обучение [Електронний ресурс] / Дьяконов А. — Режим доступа: https://alexanderdyakonov.files.wordpress.com/2017/06/book_boosting_pdf.pdf

22. Sokolova M. Beyond Accuracy, F-score and ROC: a Family of Discriminant
Measures for Performance Evaluation [Електронний ресурс] / Sokolova

M., Japkowicz N., Szpakowicz S. – American Association for Artificial Intelligence — Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.549.2795&rep=rep1&type=pdf

23. Machine Learning Repository [Електронний ресурс] — Режим доступа: http://archive.ics.uci.edu/ml/datasets/extention+of+ZAlizadeh+sani+dataset

# ДОДАТОК А

## Лістинг програми

```
import pandas as pd
import numpy as np

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
roc_auc_score, precision_recall_curve,log_loss
from sklearn.model_selection import train_test_split, KFold
from sklearn.preprocessing import MinMaxScaler,Normalizer, scale
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectFromModel, RFECV, SelectKBest,
mutual_info_classif
from sklearn.decomposition import PCA
from sklearn.ensemble import ExtraTreesClassifier,RandomForestClassifier
from xgboost import XGBClassifier
from vecstack import stacking
from sklearn.naive_bayes import GaussianNB,BernoulliNB
from sklearn.tree import DecisionTreeClassifier

df = pd.read_excel('CAD.xlsx')
df.head()
df.columns
Y_LAD = df['LAD'].replace({'Stenotic':1, 'Normal':0})
Y_LCX = df['LCX'].replace({'Stenotic':1, 'Normal':0})
Y_RCA = df['RCA'].replace({'Stenotic':1, 'Normal':0})
df = df.drop(labels = ['LAD','LCX','RCA'],axis = 1)
df['Sex'] = df['Sex'].replace({"Male":1,"Fmale":0})
df['Cath'] = df['Cath'].replace({"CAD":1,"Normal":0})
df['VHD'] = df['VHD'].replace({'mild':1,'Severe':2, 'Moderate':3, 'N':0})
df['BBB'] = df['BBB'].replace({'LBBB':1, 'RBBB':2})
pd.value_counts(df['Cath'])
#replacing all Y/N values in columns
def replace_Y_N(row):
for col in df.columns:
if row[col] == "Y":
row[col] = 1
if row[col] == "N":
row[col] = 0
return row
df = df.apply(replace_Y_N,axis = 1)
for col in df.columns:
```

```
print (col)
print (df[col].unique())
df['K'].hist()
x = df.iloc[:,:-1]
y = df.iloc[:,-1]
import scipy
scipy.sparse.issparse(x.as_matrix())
def split_data(x,y):
X_train,X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.2, random_state=42)
Y_test = Y_test.reset_index()['Cath']
X_test = X_test.reset_index().iloc[:,1:]
return X_train,X_test,Y_train, Y_test
X_train,X_test,Y_train, Y_test = split_data(x,y)
pd.value_counts(Y_train)
pd.value_counts(Y_test)


from sklearn.utils import class_weight
class_weight = class_weight.compute_class_weight('balanced', np.unique(Y_train), Y_train)
class_weight
weights = {1:0.69942197,0:1.75362319}
# # Feature Selection
#2. features selected by Decision trees
clf = DecisionTreeClassifier(random_state=241,splitter = 'random',class_weight=weights,
max_depth=20)
clf.fit(X_train,Y_train)
model = SelectFromModel(clf, prefit=True)
X_DT = model.transform(x)
X_DT_train = model.transform(X_train)
X_DT_test = model.transform(X_test)
#3. RFECV
clf = LogisticRegression(random_state=241,class_weight=weights)
selector = RFECV(estimator=clf, cv=5,scoring='neg_mean_squared_error')
selector.fit(X_train, Y_train)
print('Optimal number of features: %d' % selector.n_features_)
dict(zip(list(df.columns),selector.support_))
x_rfecv = selector.transform(x)
X_train_rfecv = selector.transform(X_train)
X_test_rfecv = selector.transform(X_test)
from sklearn.utils import class_weight
class_weight = class_weight.compute_class_weight('balanced', np.unique(Y_train), Y_train)
class_weight
weights = {1:0.69942197,0:1.75362319}
# # Feature Selection
```

```
#2. features selected by Decision trees
clf = DecisionTreeClassifier(random_state=241,splitter = 'random',class_weight=weights,
max_depth=20)
clf.fit(X_train,Y_train)
model = SelectFromModel(clf, prefit=True)
X_DT = model.transform(x)
X_DT_train = model.transform(X_train)
X_DT_test = model.transform(X_test)
#3. RFECV
clf = LogisticRegression(random_state=241,class_weight=weights)
selector = RFECV(estimator=clf, cv=5,scoring='neg_mean_squared_error')
selector.fit(X_train, Y_train)
print('Optimal number of features: %d' % selector.n_features_)
dict(zip(list(df.columns),selector.support_))
x_rfecv = selector.transform(x)
X_train_rfecv = selector.transform(X_train)
X_test_rfecv = selector.transform(X_test)
selector = RFECV(estimator=clf, cv=5,scoring='neg_mean_squared_error')
selector.fit(train, Y_train)
train1 = selector.transform(train)
test1 = selector.transform(test)
if returnx is True:
x1 = selector.transform(x)
if how == 'f_classif':
selector = SelectKBest(k=30)
fitted = selector.fit(train, Y_train)
train1 = fitted.transform(train)
test1 = fitted.transform(test)
if returnx is True:
x1 = fitted.transform(x)
if how == 'mutual_info':
selector = SelectKBest(mutual_info_classif, k=30)
fitted = selector.fit(train, Y_train)
train1 = fitted.transform(train)
test1 = fitted.transform(test)
if returnx is True:
x1 = fitted.transform(x)
if how == 'svm':
lsvc = LinearSVC(C=0.1, penalty="l1", dual=False).fit(train, Y_train)
model = SelectFromModel(lsvc, prefit=True)
train1 = model.transform(train)
test1 = model.transform (test)
if returnx is True:
x1 = model.transform(x)
```

```
if how == 'pca':
pca = PCA(n_components=30)
pca.fit(train)
train1 = pca.transform(train)
test1 = pca.transform(test)
if returnx is True:
x1 = pca.transform(x)
if returnx is True:
return train1,test1, x1
else:
return train1,test1
fs_list = ['decision_trees','rfecv','f_classif','mutual_info','svm','pca']
# # ----------------------------------------
# rescaling data
def rescaling (train ,test, how):
if how == 'minmax':
scaler = MinMaxScaler(feature_range=(0, 1))
train1 = scaler.fit_transform(train)
test1 = scaler.transform(test)
if how == 'stand':
selector = RFECV(estimator=clf, cv=5,scoring='neg_mean_squared_error')
selector.fit(train, Y_train)
train1 = selector.transform(train)
test1 = selector.transform(test)
if returnx is True:
x1 = selector.transform(x)
if how == 'f_classif':
selector = SelectKBest(k=30)
fitted = selector.fit(train, Y_train)
train1 = fitted.transform(train)
test1 = fitted.transform(test)
if returnx is True:
x1 = fitted.transform(x)
if how == 'mutual_info':
selector = SelectKBest(mutual_info_classif, k=30)
fitted = selector.fit(train, Y_train)
train1 = fitted.transform(train)
test1 = fitted.transform(test)
if returnx is True:
x1 = fitted.transform(x)
if how == 'svm':
lsvc = LinearSVC(C=0.1, penalty="l1", dual=False).fit(train, Y_train)
model = SelectFromModel(lsvc, prefit=True)
train1 = model.transform(train)
```

```
test1 = model.transform (test)
if returnx is True:
x1 = model.transform(x)
if how == 'pca':
pca = PCA(n_components=30)
pca.fit(train)
train1 = pca.transform(train)
test1 = pca.transform(test)
if returnx is True:
x1 = pca.transform(x)
if returnx is True:
return train1,test1, x1
else:
return train1,test1
fs_list = ['decision_trees','rfecv','f_classif','mutual_info','svm','pca']
# # ----------------------------------------
# rescaling data
def rescaling (train ,test, how):
if how == 'minmax':
scaler = MinMaxScaler(feature_range=(0, 1))
train1 = scaler.fit_transform(train)
test1 = scaler.transform(test)
if how == 'stand':
### Logistic regression
#clf = LogisticRegression(random_state=241,class_weight={1:0.4,0:0.6})
#clf = LogisticRegression(random_state=241)
clf = LogisticRegression(random_state=241,class_weight=weights)
clf.fit(X_train,Y_train)
res = clf.predict(X_test)
estimate_model(res)
res1 = clf.predict_proba(X_test)
log_loss(Y_test,res1)
#### rescaled
clf = LogisticRegression(random_state=241,class_weight=weights)
for s in sc:
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
clf.fit(X_train1,Y_train)
res = clf.predict(X_test1)
estimate_model(res)
print ('\n\n')
#### after feauture selection
clf = LogisticRegression(random_state=241,class_weight=weights)
for i in fs_list:
```

```
print ('--------',i,'--------')
train,test = feature_selection(i)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
#rescaling + feature selection
clf = LogisticRegression(random_state=241,class_weight=weights)
for i in fs_list:
for s in sc:
print ('--------',i,'--------')
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
train,test = feature_selection(i,X_train1, X_test1)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
# ### Naive Bayes Classifier
# #### Gaussian
clf = GaussianNB()
clf.fit(X_train,Y_train)
res = clf.predict(X_test)
estimate_model(res)
# #### Bernoulli
clf = BernoulliNB()
clf.fit(X_train,Y_train)
res= clf.predict(X_test)
estimate_model(res)
clf = BernoulliNB()
for s in sc:
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
clf.fit(X_train1,Y_train)
res = clf.predict(X_test1)
estimate_model(res)
print ('\n\n')
# ### after feauture selection
clf = BernoulliNB()
for i in fs_list:
print (i)
train,test = feature_selection(i)
clf.fit(train,Y_train)
res = clf.predict(test)
```

```
estimate_model(res)
print ('\n\n')
#rescaling + feature selection
clf = BernoulliNB()
for i in fs_list:
for s in sc:
print (i)
print (s)
X_train1,X_test1 = rescaling (X_train, X_test, s)
train,test = feature_selection(i,X_train1, X_test1)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
# #### Decision Trees
clf = DecisionTreeClassifier(random_state=241,splitter = 'random',class_weight=weights,
max_depth=20)
clf.fit(X_train,Y_train)
res = clf.predict(X_test)
estimate_model(res)
clf.feature_importances_
for i in range(len(clf.feature_importances_)):
if clf.feature_importances_[i]>0:
print (df.columns[i],clf.feature_importances_[i] )
# ### after feauture selection
clf = DecisionTreeClassifier(random_state=241,splitter = 'random')
for i in fs_list:
print ('--------',i,'--------')
train,test = feature_selection(i)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
clf = DecisionTreeClassifier(random_state=241,splitter = 'random')
for i in fs_list:
for s in sc:
print ('--------',i,'--------')
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
train,test = feature_selection(i,X_train1, X_test1)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
```

```
# ## LDA
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
clf = LinearDiscriminantAnalysis()
clf.fit(X_train, Y_train)
clf = LinearDiscriminantAnalysis()
for i in fs_list:
for s in sc:
print ('--------',i,'--------')
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
train,test = feature_selection(i,X_train1, X_test1)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
# ## Multi-layer Perceptron classifier.
from sklearn.neural_network import MLPClassifier
X_train1,X_test1 = rescaling (X_train, X_test, 'stand')
train,test = feature_selection('decision_trees',X_train1, X_test1)
clf = MLPClassifier(activation='tanh',learning_rate='adaptive',alpha =0.01, max_iter=200)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
clf = MLPClassifier(activation='tanh',learning_rate='adaptive',alpha =0.01, max_iter=200)
for i in fs_list:
for s in sc:
print ('--------',i,'--------')
print ('--------',s,'--------')
X_train1,X_test1 = rescaling (X_train, X_test, s)
train,test = feature_selection(i,X_train1, X_test1)
clf.fit(train,Y_train)
res = clf.predict(test)
estimate_model(res)
print ('\n\n')
```