НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО» КАФЕДРА ІНФОРМАТИКИ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

**КУРСОВА РОБОТА**

з дисципліни «Аналіз даних в інформаційних системах»

на тему: «Прогнозування наявності серцево-судинного захворювання у людини на основі медичних показників. Методи K-Nearest Neighbors, Logistic Regression, Random Forest, SVM»

Студента 2 курсу групи ІП-11 Спеціальності: 121

«Інженерія програмного забезпечення» Панченка Сергія Віталійовича

«ПРИЙНЯВ» з оцінкою

доц. Ліхоузова Т.А. / доц. Олійник Ю.О.

Підпис Дата

Київ - 2023 рік

Національний технічний університет України “КПІ ім. Ігоря Сікорського”

Кафедра інформатики та програмної інженерії

Дисципліна Аналіз даних в інформаційно-управляючих системах Спеціальність 121 "Інженерія програмного забезпечення"

Курс 2 Група ІП-11 Семестр 4

**ЗАВДАННЯ**

**на курсову роботу студента**

Панченка Сергія Віталійовича

|  |  |
| --- | --- |
| 1.Тема роботи | Прогнозування наявності серцево-судинного захворювання у людини на |
| основі медичних показників. Методи K-Nearest Neighbors, Logistic Regression, Random Forest, SVM. | |

|  |  |
| --- | --- |
| 2.Строк здачі студентом закінченої роботи | 15.06.2022 |

|  |  |
| --- | --- |
| 3. Вхідні дані до роботи | методичні вказівки до курсової роботи, обрані дані з сайту |
| https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset | |

1. Зміст розрахунково-пояснювальної записки (перелік питань, які підлягають розробці)

|  |
| --- |
| Вступ, постановка задачі, аналіз предметної області, робота з даними, інтелектуальний |
| аналіз, висновки, перелік посилань, додаток А. |

1. Перелік графічного матеріалу ( з точним зазначенням обов’язкових креслень )

|  |  |
| --- | --- |
| 6.Дата видачі завдання | 16.03.2023 |

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Назва етапів курсової роботи | Термін  виконання етапів роботи | Підписи  керівника, студента |
| 1. | Отримання теми курсової роботи | 16.03.2023 |  |
| 2. | Визначення зовнішніх джерел даних | 16.03.2023 |  |
| 3. | Пошук та вивчення літератури з питань курсової роботи | 16.03.2023 |  |
| 4. | Обробка та аналіз даних | 16.03.2023 |  |
| 5. | Обґрунтування методів інтелектуального аналізу даних | 16.03.2023 |  |
| 6. | Застосування та порівняння ефективності методів  інтелектуального аналізу даних | 16.03.2023 |  |
| 7. | Підготовка пояснювальної записки | 16.03.2023 |  |
| 8. | Здача курсової роботи на перевірку | 16.03.2023 |  |
| 9. | Захист курсової роботи | 16.03.2023 |  |

|  |  |  |
| --- | --- | --- |
| Студент |  | Панченко С. В. |
| (підпис) |  | (прізвище, ім’я, по батькові) |

|  |  |  |
| --- | --- | --- |
| Керівник |  | доц. Ліхоузова Т.А |
| (підпис) |  | (прізвище, ім’я, по батькові) |
| Керівник |  | доц. Олійник Ю.О. |
| (підпис) |  | (прізвище, ім’я, по батькові) |

"26" червня 2023 р.

**АНОТАЦІЯ**

Пояснювальна записка до курсової роботи: 21 сторінки, 16 рисунків, 10 посилань.

Об’єкт дослідження: інтелектуальний аналіз даних.

Предмет дослідження: створення програмного забезпечення, що проводить аналіз даних з подальшим прогнозуванням та графічним відображенням результатів.

Мета роботи: пошук, аналіз та обробка даних, реалізація ПЗ, що використовує отримані дані для подальшого аналізу та прогнозування результату.

Дана курсова робота включає в себе: постановку задачі, аналіз предметної області, роботу з даними, аналіз обраних методів для прогнозування та їх порівняння.

ДАТАСЕТ, ПРОГНОЗУВАННЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, K-NEAREST NEIGHBORS, DECISION TREE CLASSIFIER.

**Table of Contents**

[ВСТУП 6](#__RefHeading___Toc5101_4033280663)

[1 ПОСТАНОВКА ЗАДАЧІ 7](#__RefHeading___Toc5103_4033280663)

[2 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ 8](#__RefHeading___Toc5105_4033280663)

## ВСТУП

Зараз все більше стає запитаною медична діагностика різних захворювань, що є дуже важливим аспектом для лікарів та здоров’я пацієнтів. Аналізуючи захворювання, особливо хочеться виділити цукровий діабет, адже з розвитком рівня життя цукровий діабет все частіше зустрічається в повсякденні. Саме через це швидкість та точність діагностування цукрового діабету – завдання, що потребує швидких та точних результатів.

У рамках даної курсової роботи були проаналізовані дані лікарень з різними важливими характеристиками для визначення цукрового діабету та на основі отриманих даних було використано декілька методів для прогнозування даного захворювання.

Дана курсова робота буде розроблена з використанням технологій та бібліотек Python 3[1], Pandas[2], Seaborn[3], Matplotlib[4], Sklearn[5], NumPy[6].

## 1 ПОСТАНОВКА ЗАДАЧІ

Виконання даної курсової роботи потребує виконання декількох задач: аналізу предметної області; роботи з датасетом: завантаження, дослідження його структури та виправлення наявних помилок; вибір методів для прогнозування та обґрунтування даного вибору; аналіз отриманих результатів кожного з методів та порівняння отриманих результатів ефективності.

Створення застосунку, що поділяє отримані дані на тренувальні та тестові для перевірки декількох методів, у подальшому порівняння ефективності методів. Для прогнозування буде використано методи K-Nearest Neighbors, Logistic Regression, Random Forest, SVM. Для кожного методу проаналізувати результати та в кінці порівняти результати цих методів. Обрати найоптимальніший метод для прогнозування хвороби.

Виконане дослідження можна буде використовувати для прогнозування можливої серцево-судинної хвороби.

Вхідними даними будуть вік, зріст, вага, стать, систолічний кров’яний тиск, діастолічний артеріальний тиск, кількість холестеролу в крові, кількість глюкози в крові, чи курить пацієнт, чи вживає алкоголь, чи займається фізичною активністю.

## 2 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Серцево-судинні захворювання (ССЗ) є одними з найбільш поширених причин смерті в світі. ССЗ можуть включати такі захворювання, як ішемічна хвороба серця (ІХС), цереброваскулярні захворювання (ЦВЗ), артеріальна гіпертензія (АГ), хронічна обструктивна хвороба легень (ХОЛЛ), діабет та інші.

Статистика ССЗ може відрізнятися залежно від країни та регіону, а також від факторів ризику, таких як вік, стать, стиль життя і генетика. Однак загалом ССЗ становлять значну частку смертності в світі.

За даними Всесвітньої організації охорони здоров'я (ВООЗ), в 2019 році ССЗ стали причиною смерті для більше ніж 17 мільйонів людей, що становить більше 30% від загальної кількості смертей в світі. Від ІХС та ЦВЗ померли понад 13 мільйонів людей, що становить більше 80% від усіх смертей від ССЗ. Більшість смертей від ССЗ відбуваються в країнах з низьким і середнім рівнем доходів.

Науковці вважають, що багато випадків ССЗ можна запобігти шляхом зменшення факторів ризику, таких як неправильне харчування, недостатній рівень фізичної активності, куріння та надмірне вживання алкоголю. Зменшення факторів ризику та поліпшення доступу до належної медичної допомоги може допомогти знизити рівень ССЗ та покращити здоров'я людей.

У програмному забезпеченні буде реалізовано наступну функціональність, що включає в себе:

* завантаження та дослідження структури датасету;
* використання декількох моделей прогнозування даних;
* прогнозування за характеристиками наявність серцево-судинного захворювання у людини;
* відображення отриманих результатів та їх аналіз;
* порівняння використаних методів.