Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

бакалавр
(назва освітнього ступеня)

на тему: Створення інформаційної системи для аналізу медичних даних

Виконав: студент IV курсу, групи СНс-42
спеціальності 122 Комп’ютерні науки
(шифр і назва спеціальності)

_____ Дацко М.І.
(підпис) (прізвище та ініціали)

Керівник _____ Гащин Н.Б.
(підпис) (прізвище та ініціали)

Нормоконтроль _____ Шимчук Г.В.
(підпис) (прізвище та ініціали)

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент _____ _____
(підпис) (прізвище та ініціали)

Тернопіль - 2024

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«__» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Бакалавр
(назва освітнього ступеня)

за спеціальністю 122 Комп’ютерні науки
(шифр і назва спеціальності)

Студенту Дацку Мар’яну Ігоровичу
(прізвище, ім’я, по батькові)

1. Тема роботи Створення інформаційної системи для аналізу медичних даних

Керівник роботи Гащин Надія Богданівна, к.т.н., доц.
(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «29» 04 2024 року № 4/7-472

2. Термін подання студентом завершеної роботи 27.06.2024 р.

3. Вихідні дані до роботи наукові літературні джерела

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ
1. Специфіка даних, що вивчаються, та методи обробки, що застосовуються в роботі.
2. Побудова моделей та людино-машинного інтерфейсу
3. Технічна реалізація системи
4. Безпека життєдіяльності, основи охорони праці
Висновки

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень, слайдів)
1. Титулка. 2. Актуальність. 3. Мета, задачі дослідження.
4. Характеристики вхідних медичних даних. 5. Метод групового урахування аргументів.
6. Схема МГУА. 7. Програмні засоби розробки. 8. Схема алгоритму визначення моделі-переможця. 9. Алгоритм розрахунку ризику розвитку АГ (артеріальної гіпертонії).
10. Архітектура системи. 11. Скріншоти роботи створеного ПЗ
12. Висновки. Основні результати проведеного дослідження

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Безпека життєдіяльності, основи хорони праці			

7. Дата видачі завдання 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	29.04 – 30.04.24	*Виконано*
2.	Підбір джерел про інформаційні моделі та системи		
	медичного характеру	01.05 – 05.05.24	*Виконано*
3.	Опрацювання джерел про моделі та інформаційні		
	системи аналізу медичних даних	06.05 – 11.05.24	*Виконано*
4.	Проведення дослідження щодо побудови моделей та		
	Створення людино-машинного інтерфейсу	12.05 – 15.05.24	*Виконано*
5	Розроблення програмного коду	15.05 – 18.05.24	*Виконано*
6.	Оформлення розділу «Специфіка даних, що вивчаються,		
	та методи обробки, що застосовуються в роботі»	19.05 – 22.05.24	*Виконано*
7.	Оформлення розділу «Побудова моделей та людино-		
	машинного інтерфейсу»	23.05 – 27.05.24	*Виконано*
8.	Оформлення розділу «Технічна реалізація системи»	28.05 – 01.06.24	*Виконано*
9.	Виконання завдання до підрозділу «Безпека		
	життєдіяльності, основи хорони праці»	02.06 – 04.06.24	*Виконано*
10.	Оформлення кваліфікаційної роботи	04.06 – 08.06.24	*Виконано*
11.	Нормоконтроль	05.06 – 09.06.24	*Виконано*
12.	Перевірка на плагіат	10.06 – 14.06.24	*Виконано*
13.	Попередній захист кваліфікаційної роботи	14.06 – 16.06.24	*Виконано*
14.	Захист кваліфікаційної роботи	28.06.24	

Студент _____ Дацко М.І..
(підпис) (прізвище та ініціали)

Керівник роботи _____ Гащин Н.Б.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Створення інформаційної системи для аналізу медичних даних // Дацко Мар’ян Ігорович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп’ютерно-інформаційних систем та програмної інженерії, кафедра комп’ютерних наук, група СНс–42 // Тернопіль, 2024 // С. – 54, рис. – 18, табл. –2, слайдів – 12, бібліогр. – 26.

Ключові слова: інформаційна система, машинне навчання, регресійна модель, мультиколінеарність, коефіцієнт детермінації, передобробка даних

Кваліфікаційна робота присвячена побудові інформаційної системи для аналізу та прогнозування ризиків розвитку артеріальної гіпертонії із використанням машинного навчання.

Здійснено постановку задачі із вказанням усіх необхідних параметрів та характеристик. Проведена передобробка даних, отриманих від працівників медичних установ. Побудовано та проаналізовано спеціалізовані моделі для врахування медичних факторів ризику виникнення захворювання. Здійснено вибір потрібних критеріїв для селекції моделей. Із використанням математичного апарату індуктивного моделювання багатопараметричних даних обґрунтовано якісний вибір моделі-переможця. Розроблено алгоритм прогнозування ризику розвитку серцево-судинного захворювання із застосуванням можливостей машинного навчання. Створено довірчий інтервал для забезпечення прогнозування.

Побудовано архітектуру програмного представлення системи. Розроблено програмний інструмент для застосування системи. Роботу перевірено на наборі тестових даних. Результати тестування системи свідчать про високі результати прогнозування.

ANNOTATION

Creation of an information system for the analysis of medical data // Datsko Marian // Ternopil Ivan Pul'uj National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science // Ternopil, 2024 // P. - 54, Fig. - 18, Table - 2, Slide - 12, References - 26.

Keywords: information system, machine learning, regression model, multicollinearity, coefficient of determination, data preprocessing

Thesis deals with the construction of an information system for the analysis and forecasting of the risks of the development of arterial hypertension using machine learning.

The task was formulated with all necessary parameters and characteristics specified. Processing of data received from employees of medical institutions has been carried out. Specialized models were built and analyzed to take into account medical risk factors for the occurrence of the disease. The selection of the necessary criteria for the selection of models was made. Using the mathematical apparatus of inductive modeling of multiparametric data, the qualitative selection of the winning model is justified. An algorithm for predicting the risk of developing cardiovascular disease using machine learning capabilities has been developed. A confidence interval was created to support the prediction.

The architecture of the software representation of the system has been built. A software tool for system application has been developed. The work is verified on a set of test data. The results of system testing indicate high forecasting results.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ СКОРОЧЕНЬ І ТЕРМІНІВ

API (Application Programming Interface) - набір певних правил і протоколів, які дозволяють різним програмам взаємодіяти одна з одною.

BMI (Body Mass Index) – індекс маси тіла.

REST (Representational State Transfer) – архітектурний стиль, який є набором принципів та обмежень, які визначають, як клієнти та сервери повинні взаємодіяти між собою.

SCORE – Systematic COronary Risk Evaluation.

АГ (артеріальна гіпертонія) – хронічне захворювання, яке обумовлено підвищенням артеріального тиску систоли або артеріального діастолічного тиску в стані спокою.

БД – база даних.

ІС (інформаційна система) – сукупність організаційних і технічних засобів для збереження та обробки інформації з метою забезпечення інформаційних потреб користувачів.

МГУА (метод групового урахування аргументів) – сімейство індуктивних алгоритмів для математичного моделювання багатопараметричних даних.

МН (машинне навчання) – розділ штучного інтелекту, який досліджує методи, що дозволяють комп'ютерам покращувати свої характеристики на основі отриманого досвіду.

МНК (метод найменших квадратів) – метод знаходження наближеного розв'язку надлишково-визначеної системи.

ССЗ – серцево-судинне захворювання.

СУБД – система управління базами даних.

Шкала SCORE – розроблена для оцінки ризику смертельного серцево-судинного захворювання упродовж 10 років.

ЗМІСТ

ВСТУП

Актуальність теми. Зі зростанням чисельності населення та збільшенням кількості промислових міст дедалі більше людей страждають від погіршення екології. Загальний стан здоров'я людей падає у зв'язку із забрудненням довкілля.

Паралельно з цим, науково-технологічний прогрес у галузі медицини дозволяє діагностувати та шукати методи лікування для багатьох тяжких захворювань. Тим не менш, необхідні апарати для діагностики та різні лікарські засоби досі, здебільшого, існують лише у великих містах.

Беручи до уваги згадане вище, виділяють кілька проблем: погіршення стану здоров'я населення, що говорить про необхідність збільшення кількості досліджень, а також про нестачу необхідного обладнання в сільських місцевостях. Для вирішення цих проблем потрібне дослідження, яке дозволить виявити відмінності у станах здоров'я населення міста та села з метою збільшення фінансування сільських амбулаторій.

ССЗ, а саме АГ є одним із найпоширеніших хронічних захворювань у Україні. Низька ефективність лікування та смертність цієї хвороби визначає актуальність проблеми. Поширеність АГ у Україні за даними [1] становить більше 40%, і це число невпинно зростає. При високій поширеності цієї хвороби ефективність лікування та контроль факторів виникнення залишаються в Україні на досить низькому рівні. Одним із найвідоміших інструментів прогнозування ризику розвитку ССЗ є шкала SCORE [1].

Мета роботи – автоматизація процесу створення моделей за допомогою створення застосунку, що забезпечує опис залежностей у медичних даних та прогнозування ризику розвитку АГ.

Для досягнення мети виділено ряд завдань:

- попередньо обробити дані;
- підібрати моделі;
- вибрати необхідні критерії для селекції моделей;
- створити інструменти для побудови та перевірки якості моделей;

– вибрати єдину модель-переможця;
– застосувати тестові дані до побудованої моделі;
– створити інтерфейс для використання збудованої моделі;
– створити довірчий інтервал для прогнозованого значення;
– проаналізувати результати та сформулювати висновки;
– окреслити перспективи розвитку системи.

Практична цінність роботи. Створена ІС, що використовує метод МН для побудови залежності у медичних даних може передбачати ризик розвитку ССЗ, тому її можна застосовувати у медичних установах.

РОЗДІЛ 1. СПЕЦИФІКА ДАНИХ, ЩО ВИВЧАЮТЬСЯ, ТА МЕТОДИ ОБРОБКИ, ЩО ЗАСТОСОВУЮТЬСЯ В РОБОТІ

1.1 Постановка задачі. Опис та характеристики даних

У рамках виконання роботи є задача побудови ІС, яка аналізує медичні показники з використанням алгоритму МН. ІС аналізує медичні дані та будує моделі, що описують залежність у цих даних, а також прогнозує ризик розвитку АГ. Дані медичного характеру, які містять показники, подані у табл. 1.1.

Таблиця 1.1 – Показники отриманих даних

Найменування показника	**Тип даних**
ПІП пацієнта	Стрічкове значення
Найменування організації	Стрічкове значення
Дата дослідження	Дата
Ризик SCORE	Числове значення, відсотки
Modification of Diet in Renal Disease (MDRD)	Числове значення, мл/хв/м^2
Рівень глюкози	Числове значення, ммоль/л
Рівень холестерину	Числове значення, ммоль/л
Систолічний тиск	Числове значення, мм. рт. ст.
Діастолічний тиск	Числове значення, мм. рт. ст.
Body Mass Index (BMI)	Числове значення, кг/м^2
Первинне/вторинне обстеження	Стрічкове значення (так/ні)
Статус паління	Стрічкове значення (так/ні/ раніше)
Стать пацієнта	Стрічкове значення (M/F)
Вік	Числове значення
Наявність серцево-судинного захворювання	Стрічкове значення (так/ні)

Перелічені в табл. 1.1 дані були отримані від медичних організацій м.

Золочів, які проводили запис вимірювань пацієнтів, які страждають на ССЗ.

Для деяких із перелічених вище показників існують коридори значень, що визначають допустимі значення вимірів. Вони представлені у табл. 1.2.

Таблиця 1.2 – Коридори значень показників

Найменування показника	Допустимі значення
Систолічний тиск	90 – 230 мм. рт. ст.
Діастолічний тиск	40 – 120 мм. рт. ст.
SCORE	0 - 50%
BMI	15 - 45
Рівень глюкози	3 - 25 ммоль/л
Рівень холестерину	2 - 15 ммоль/л
MDRD	15 - 140

1.2 Теоретичні засади МГУА

Для реалізації алгоритму МН використовувався МГУА, який заснований на переборі регресійних моделей, котрі поволі ускладнюються, і виборі кращого рішення з урахуванням зовнішнього критерію. Саме досягнення мінімуму зовнішнього критерію під час народження моделі означає, що модель є шуканою. Для породження використовуються базисні моделі, які можуть бути не тільки поліноміальні, але і нелінійні і ймовірнісні функції [2].

У більшості алгоритмів МГУА використовуються поліноміальні моделі, які можливо записати як функціональний ряд Вольтерра чи поліном Колмогорова-Габора:

$$y = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} x_i x_j x_k + \dots , \qquad (1.1)$$

де $X\ (x_1, x_2, ..., x_M)$ - вхідний вектор змінних, $A\ (a_1, a_2, ..., a_M)$ - вектор коефіцієнтів.

Саме за допомогою формули (1.1) вибирається загальний вид моделей, що перебираються, які називаються опорними функціями. З допомогою цих опорних функцій створюються різновиди моделей, котрі можуть містити як один аргумент, так і всі. Їх можна бути побудувати з однією змінною, з різними парами змінних, з різними трійками змінних і т.д. Розмірність вектора визначає складність моделі. За допомогою методів регресійного аналізу вираховуються коефіцієнти $a_1, a_2, ..., a_M$. На рис. 1.1 зображено структуру МГУА, де $x_1, x_2, ..., x_k$ – вхідні змінні, а $f(x_i, x_j)$ – залежності від кількох вхідних змінних. На наступних шарах будуються складніші залежності від побудованих функцій на попередньому шарі.

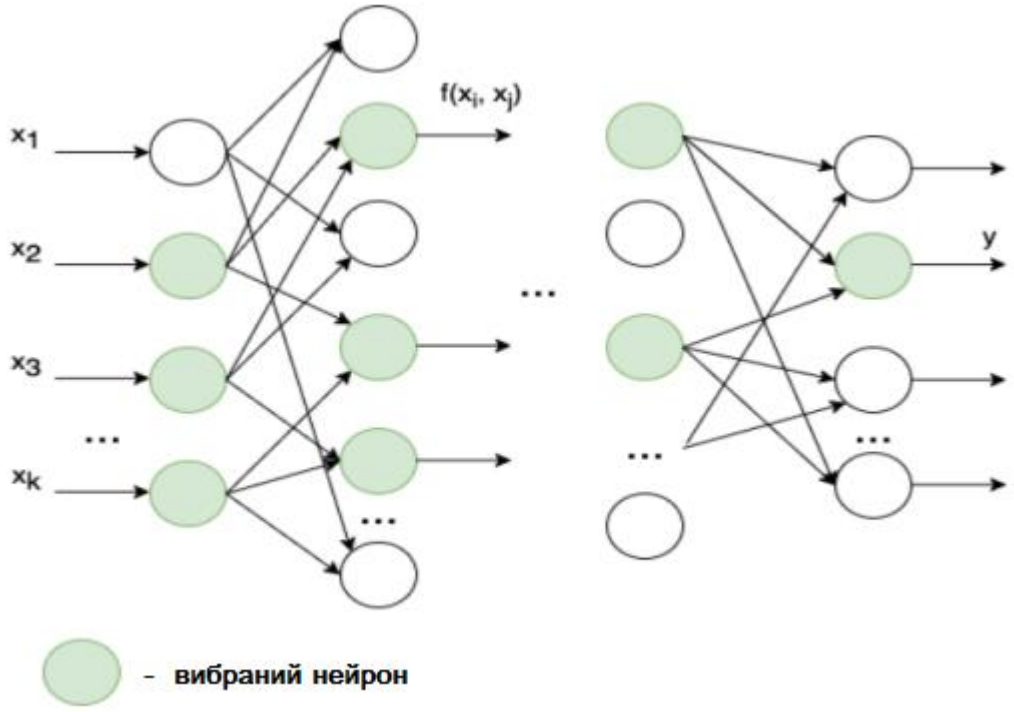

Рисунок 1.1 – Структура МГУА

Після того, як коефіцієнти моделей розраховані, серед них відбираються кілька кращих. Якість моделей встановлюється коефіцієнтом детермінації, чи середньоквадратичним відхиленням помилки, чи іншими зовнішніми критеріями [3]. Як тільки знайдено найкращу модель, з урахуванням зовнішнього критерію,

або досягнуто максимальної складності моделі, алгоритм закінчується. В іншому випадку здійснюється наступний етап алгоритму, де вхідними аргументами є моделі, побудовані на попередньому етапі.

Зазвичай вихідну вибірку ділять на дві підвибірки: навчальну та тестову. Навчальна служить до розрахунку коефіцієнтів моделі, а тестова для перевірки якості моделей. Розподіл вибірки залежить від теоретичних передумов.

Побудова моделей здійснюється при допомозі індуктивного перебірного підходу. Алгоритми МГУА відшукують єдину оптимальну модель для кожної вибірки при допомозі повного перебирання всіх моделей-кандидатів та проводять її оцінку із використанням зовнішнього критерію.

1.3 Вибір критеріїв для селекції моделей. Проблеми автокореляції, гетероскедастичності, мультиколінеарності

Після побудови рівнянь регресії, зазвичай, перевіряється загальна якість побудованих моделей, що оцінюється за тим, як добре емпіричне рівняння регресії координується із даними статистики. Мірою загальної якості створеної моделі є коефіцієнт детермінації [4], котрий показує, наскільки одна випадкова величина залежить від багатьох інших. Він обчислюється за формулою (1.2):

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma (y_i - \bar{y})^2} \qquad (1.2)$$

де $(y_i - \bar{y})^2$ - відхилення і-ї спостерігаємої точки від середнього значення $\bar{y}$ залежної змінної Y; e_i - лишки регресії.

В цьому випадку, формула показує залежність випадкової величини від факторів x. Коефіцієнт детермінації набуває значення у проміжку від 0 до 1. Неважко помітити, що якщо між x і y є значний лінійний зв'язок, тоді дисперсія помилок моделі значно менше, ніж дисперсія випадкової величини y. Отже,

коефіцієнт детермінації ближчий до 1.

Звідси, коефіцієнт детермінації дозволяє визначити наскільки добре рівняння регресії описує поведінку залежної змінної y. Можна сказати, чим зв'язок між y і x є тіснішим, тим ближчим коефіцієнт детермінації буде до 1. І навпаки.

Проте коефіцієнт детермінації, близький до 1, не гарантує високу якість побудованої моделі. До погіршення якості моделі може призвести наявність автокореляції [4]. Тому важливим критерієм селекції моделей є тестування рівнянь регресії на наявність автокореляції. Автокореляція є статистичним взаємозв'язком (кореляцію) між випадковими величинами і тими самими випадковими величинами, але взятими зі зсувом [7]. Наприклад, за часом. Зазвичай, для аналізу корелювання відхилень випадкових величин використовують статистику Дарбіна-Уотсона, яка розраховується за формулою (1.3):

$$DW = \frac{\Sigma(e_i - e_{i-1})^2}{\Sigma e_i^2}, \qquad (1.3)$$

де e_i - лишок регресійної моделі.

Статистика Дарбіна-Уотсона (DW) може бути від 0 до 4. DW = 0 говорить про позитивну автокореляцію. DW = 4 - про негативну. Якщо автокореляція відсутня, статистика DW = 2. Якщо ж кожні сусідні відхилення приблизно рівні один одному, то кожен доданок e_i - e_{i-1} наближається до 0 , а значить і сама статистика прагне нульового значення. Такий результат стверджує про лінійну залежність між лишками. У тому випадку, коли $e_i \approx e_{i-1}$ точки по черзі відхиляються в різні боки від рівняння регресії, і статистика DW прагне 4, що говорить про негативну автокореляцію лишків. Коли поведінка відхилень випадкова, можна сказати, що частина відхилень має різні знаки, інша частина має однакові знаки. Тоді статистика DW наближається до 2. Це каже, що створена модель, мабуть, відтворює реальну залежність. Такі результати

свідчать, що, швидше за все, не має неврахованих істотних факторів, котрі чинять вплив на залежну змінну.

На рис. 1.2 наведено відрізок, яким перевіряється гіпотеза щодо браку автокореляції лишків. d_l і d_U - критичні точки при рівні значимості α = 0, 01.

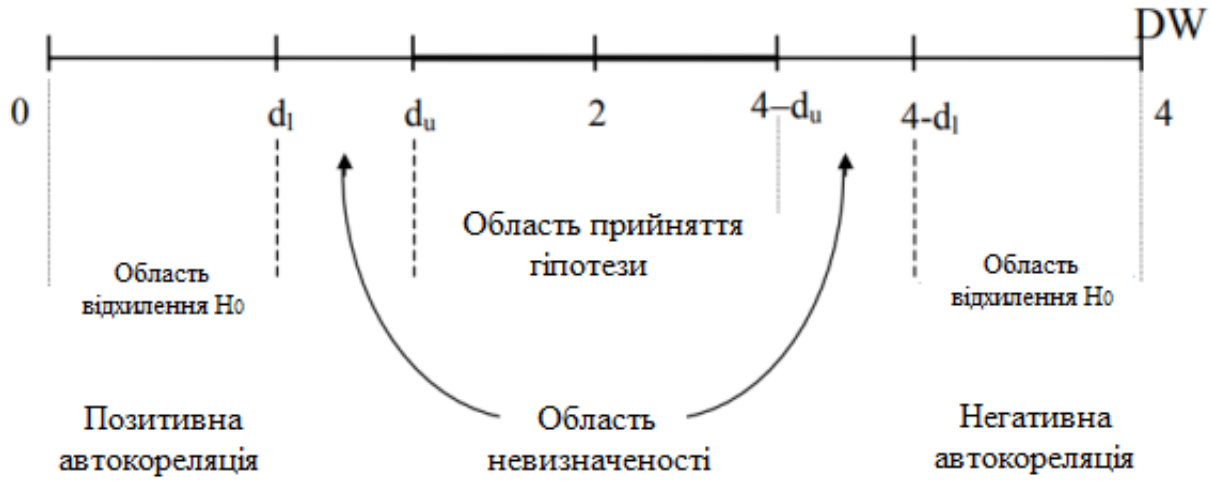

Рисунок 1.2 – Значення статистик DW

Під час проведення відбору моделей слід звернути особливу увагу на одну з найпоширеніших проблем у регресійному аналізі – гетероскедастичність [4]. Наявність гетероскедастичності говорить про неоднорідність розподілу спостережень, що виявляється у непостійній дисперсії випадкової помилки моделі. Гетероскедастичність протилежна гомоскедастичності, що означає однорідність даних (постійність дисперсії помилок) [7].

На рис. 1.3 та 1.4 показано приклад гетероскедастичності і гомоскедастичності. На рис. 1.3 дисперсія залишається однією і тією ж протягом усієї картинки, на відміну від рис. 1.4. На рис. 1.4 дисперсія не є постійною, що призводить до заниження стандартних помилок. Це свідчить, що помилки не мають однакової дисперсії, тому гетероскедастичність є проблемою.

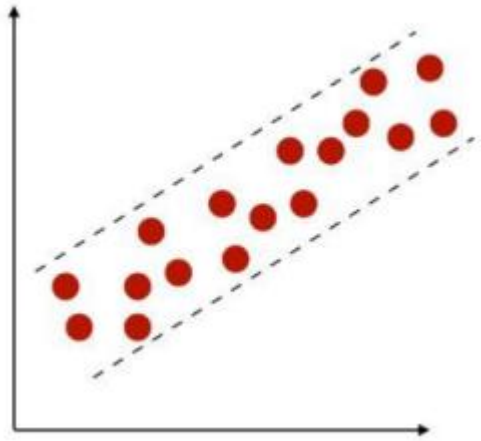

Рисунок 1.3 – Гомоскедастичність

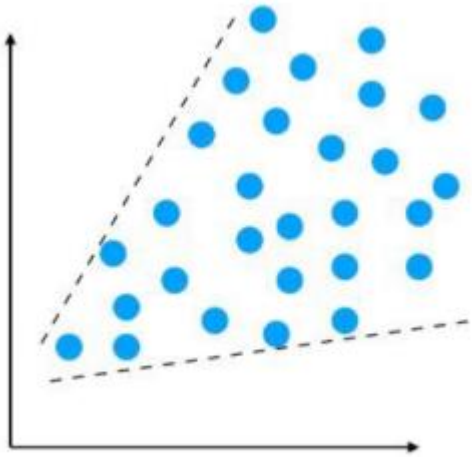

Рисунок 1.4 – Гетероскедастичність

Одним із поширених тестів на гетероскедастичність [5] є тест Бреуша-Пагана [8]. Передбачається, що дисперсія випадкової помилки залежить від кількох незалежних змінних. Тест складається з кількох етапів:

1. знаходяться лишки регресійної моделі e_i;
2. отримані лишки підносяться до квадрату - $(e_i)^2$;
3. за формулою (1.4) знаходять складову оцінку дисперсії помилок;
4. знаходять квадрати стандартизованих лишків e^2/δ^2 ;
5. оцінюють (за МНК) допоміжну лінійну регресію (формула (1.5));
6. розраховується статистика тесту за формулою (1.6);
7. тестова статистика RSS/2 має розподіл хі – квадрат із p степенями вільності;

8. порівнюються значення статистики з критичним $X^2_{\alpha}(p)$.

Якщо значення статистики перевищує критичне значення, то гіпотеза про гомоскедастичність відкидається (є гетероскедастичність). Інакше гіпотеза приймається (модель гомоскедастична).

$$\delta^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}, \tag{1.4}$$

де e_i - лишок регресійної моделі, n - обсяг вибірки.

$$\frac{e^2}{\delta^2} = \gamma_0 + \gamma_1 Z_1 + \ \dots \ + \gamma_p Z_p + u, \tag{1.5}$$

де $Z_1 \dots Z_p$ - незалежні змінні, e^2/δ^2 - квадрати стандартизованих лишків.

$$\frac{RSS}{2}, \tag{1.6}$$

де RSS - сума квадратів лишків допоміжної регресійної моделі.

Ще одна проблема, яка може виникнути при процедурі відбору моделей, є мультиколінеарність [4]. Це лінійний взаємозв'язок двох і більше змінних регресійної моделі. При цьому існує кілька її видів: повна (досконала) мультиколінеарність, часткова та висока мультиколінеарність [7]. Повна мультиколінеарність каже про функціональну залежність між пояснювальними змінними. Вона дозволяє однозначно визначити параметри регресійної моделі. Тому висновки про коефіцієнти та саму модель будуть недостатньо надійними. Така мультиколінеарність виникає вкрай рідко. Найчастіше на практиці зустрічається висока та часткова мультиколінеарність. Висока мультиколінеарність говорить про сильну кореляцію між факторами. Часткова та висока мультиколінеарність призводить до збільшення дисперсії оцінок

(стандартних помилок), і це може погіршити їхню точність.

Одним із методів виявлення мультиколлінеарності є число обумовленості матриці [6]. Можна сказати, що якщо змінні майже лінійно залежні, то визначник матриці $X^T X$ буде прагнути до 0, а елементи транспонованої матриці прагнутимуть до нескінченності. Те саме відбуватиметься з елементами матриці коваріації. Точність оцінок буде низька через великі дисперсії. Така ситуація відбувається тому, що якщо змінні пов'язані, то важко окремо оцінити вплив кожної з них. Також, варто згадати, що визначник матриці можна знайти перемноживши власні числа. Важливу інформацію про них дає індекс обумовленості матриці, що обчислюється за формулою (1.7):

$$k = \sqrt{\frac{max\,(\lambda_i)}{min\,(\lambda_i)}}, \qquad (1.7)$$

де λ_i - власні числа матриці $X^T X$.

Якщо у матриці є власні числа, близькі до 0, можна сказати про наявність мультиколінеарності. Значення k, що отрималося, свідчать про наявність колінераності наступне:

– якщо k < 10 –, відсутня;
– якщо 10 < k < 30, є невелика;
– якщо k > 30, наявна сильна.

РОЗДІЛ 2. ПОБУДОВА МОДЕЛЕЙ ТА ЛЮДИНО-МАШИННОГО ІНТЕРФЕЙСУ

2.1 Реалізація та аналіз комбінаторного алгоритму МГУА

Для створення ІС з алгоритмами МН було обрано мову програмування Python. Саме ця мова має всі необхідні бібліотеки для реалізації МГУА. При створенні системи та алгоритмів МН були використані такі бібліотеки:

– numpy, math [9] - для різних математичних обчислень;
– pandas [10] - для обробки та аналізу даних;
– itertools [11] - комбінаторний перебір варіантів;
– scikit-learn [12], statsmodel [13] - для побудови та відбору моделей, обчислення коефіцієнтів детермінації, автокореляції тощо.

Наведені вище бібліотеки добре підходять для алгоритмів МН. Вони мають усі необхідні методи реалізації складних математичних обчислень.

Одним із перших кроків у побудові ІС є попередня обробка даних. Дані були отримані не дуже добре. Наприклад, деякі значення були відсутні або значення не відповідали коридорам значень. Такі рядки було виключено з датасету. Порожні осередки прийнято виключити, а не замінювати середнім значенням. Таке рішення було прийнято, оскільки аналізуються медичні дані. Усереднені значення могли погіршити точність побудови моделі.

Також, провівши аналіз медичних джерел, було з'ясовано які фактори теоретично можуть впливати на ризик розвитку АГ. Наприклад, назва організації навряд чи впливатиме на можливість захворіти на ССЗ. Виходячи з цього, не бралися до розрахунку такі виміри як дата, найменування організації, ПІБ, первинний/вторинний огляд. Всі інші виміри були враховані у побудові залежностей:

– холестерин;
– систолічний тиск;
– діастолічний тиск;
– рівень глюкози тощо.

Дані поділялися на дві групи залежно від ризику розвитку ССЗ.

Як виявилося, дані містять не лише числові значення, а й текстові. Тому потрібно закодувати текстові значення. Для кодування використовувався модуль sklearn.preprocessing. Кодування списку з фіксованими значеннями здійснювалося за допомогою об'єкта LabelEncoder, який замінює текстові значення на 0 та 1, залежно від кількості унікальних значень у списку [14]. За допомогою цього модуля було закодовано такі стовпці: статус паління, наявність захворювання, стать. Під час кодування виникла проблема: якщо дані закодувати 0 і 1 , то при побудові логарифмічних залежностей можуть виникнути помилки, так як аргумент має бути строго більшим за 0. Тим самим було прийнято рішення закодувати текстові значення 1 і 2 .

Завантаження даних у програму здійснювалося за допомогою модуля Pandas, який містить об'єкт read_csv [10] . Ця команда дозволяє обробляти та завантажувати у програму файли у форматі csv.

Наступним етапом після попередньої обробки даних є побудова моделей. Загальний вигляд моделей можна описати за формулою (2.1):

$$Y \approx f(X_1, X_2, \dots, X_n), \tag{2.1}$$

де Y - результативний показник, $X_1, X_2, \dots, X_n$ — фактори, які можуть впливати на значення Y.

У нашому випадку, Y – ризик розвитку ССЗ, а $X_1, X_2, \dots, X_n$ – рівень холестерину, тиск, вік тощо.

На першій ітерації алгоритму було поставлено завдання побудувати моделі із формули (2.2):

$$
\begin{aligned}
&Y \approx b_0 + b_1 X_i + \varepsilon, \\
&Y \approx b_0 + b_1 X_i + b_2 X_j + \varepsilon, \\
&\dots \\
&Y \approx b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + \varepsilon, \\
&Y \approx b_0 + b_1 X_i + b_2 X_i^2 + \varepsilon, \\
&Y \approx b_0 + b_1 X_i^2 + \varepsilon, \\
&Y \approx b_0 + b_1 X_i^2 + b_2 X_j^2 + \varepsilon, \\
&\dots \\
&Y \approx b_0 + b_1 X_1^2 + b_2 X_2^2 + \dots + b_n X_n^2 + \varepsilon, \\
&Y \approx b_0 + b_1 ln\, X_i + \varepsilon, \\
&Y \approx b_0 + b_1 ln\, X_i + b_2 ln\, X_j + \varepsilon, \\
&\dots \\
&Y \approx b_0 + b_1 ln\, X_1 + b_2 ln\, X_2 + \dots + b_n ln\, X_n + \varepsilon.
\end{aligned}
\tag{2.2}
$$

де Y - результативний показник, $X_1, X_2, \dots, X_n$ - пояснюючі фактори, $b_1, b_2, \dots, b_n$ - коефіцієнти рівнянь, ε - випадкова величина.

Після обробки даних було відібрано 9 факторів, що пояснюють:

- рівень глюкози;
- рівень холестерину;
- систолічний артеріальний тиск;
- діастолічний артеріальний тиск;
- BMI;

– статус куріння;
– стать;
– вік;
– наявність ССЗ.

Як можна помітити, моделі з формули (2.2) можна розділити на 4 групи:

– лінійні;
– логарифмічні;
– квадратичні;
– змішані.

Практично у всіх моделей із цих груп можна виділити таку подібність: кількість доданків збільшується від 2 до n+1. При цьому потрібно враховувати всілякі комбінації факторів, що пояснюють. Наприклад, лінійна модель, що складається з 4 доданків з урахуванням вільного члена, включає ще множину моделей. У нашому випадку пояснюючих факторів 9. Тому отримаємо множину моделей, що містить всі комбінації по 3 з 9 пояснюючих факторів без повторень. Кількість варіантів можна одержати, використовуючи формулу поєднань без повторень. Так як практично в кожній групі моделей кількість доданків варіюється від 2 до n+1, можна зробити висновок, що слід використовувати формулу поєднань, створюючи комбінації по 1,2,...,n з 9 факторів, що пояснюють.

Для створення поєднань без повторень використовувалася бібліотека itertools. В якій є метод комбінацій, що приймає на вхід масив об'єктів, з яких будуються комбінації, та довжину комбінацій [11]. Тому для кожної моделі з 4 груп були реалізовані методи, які будують різні варіанти з масиву чисел від 0 до 8 (індексація в Python починається з 0) довжиною від 1 до 9. Наприклад, при побудові моделі з 4 доданками були створені всілякі комбінації по 3 із масиву чисел від 0 до 8 ((0, 1, 2), (0, 1, 3), (0, 1, 4), ...). І для кожного варіанта були побудовані моделі, які включали саме ті стовпці з датасету, що входять до кожної комбінації (модель зі стовпцями за номером 0, 1,2; модель зі стовпцями за номером 0, 1, 3; ...). Таким чином, алгоритм породжує множину моделей, які надалі проходять відбір за зовнішніми критеріями.

Після відбору необхідних стовпців з датасета здійснювалося з'єднання у один глобальний масив у межах кожної комбінації. Залежно від кожної групи моделей здійснювалися необхідні дії із створеним масивом елементів. Наприклад, для моделей з логарифмічної групи кожного елемента зі створеного масиву брався натуральний логарифм. Для створення моделей із квадратичної групи кожен елемент масиву зводився у квадрат.

Наступним етапом реалізації алгоритму є знаходження коефіцієнтів для моделі лінійної регресії. Для цього використовувалася бібліотека sklearn.liner_model, яка допомагає розв'язувати систему рівнянь МНК [15]. Ця бібліотека містить об'єкт fit, який приймає на вхід масив X (тестові дані) та масив Y (залежні змінні) [12]. За допомогою цього об'єкта здійснюється розрахунок параметрів зазначених даних.

Таким чином, моделі з формули (9) були поділені на 4 групи:

– лінійна;

– квадратична (кожний доданок містить число з вибірки, піднесене до квадрату);

– логарифмічна (кожний доданок містить число з вибірки, прологарифмоване за експонентою);

– змішана (один із доданків містить число без змін, інший піднесений до квадрату).

На кожну групу було реалізовано окремий метод, який породжував сімейство моделей за допомогою комбінаторних обчислень. Також, у даному методі у кожної з породженої моделі обчислювалися коефіцієнти МНК. Загальна схема алгоритму МГУА представлена на рис. 2.1.

Також варто відзначити, що досить було побудувати тільки сімейство моделей з формули (2,2). Це зумовлено хорошими показниками зовнішніх критеріїв. Інакше виникла б необхідність будувати інші сімейства, поки не знайдуться моделі, що показують хороші результати.

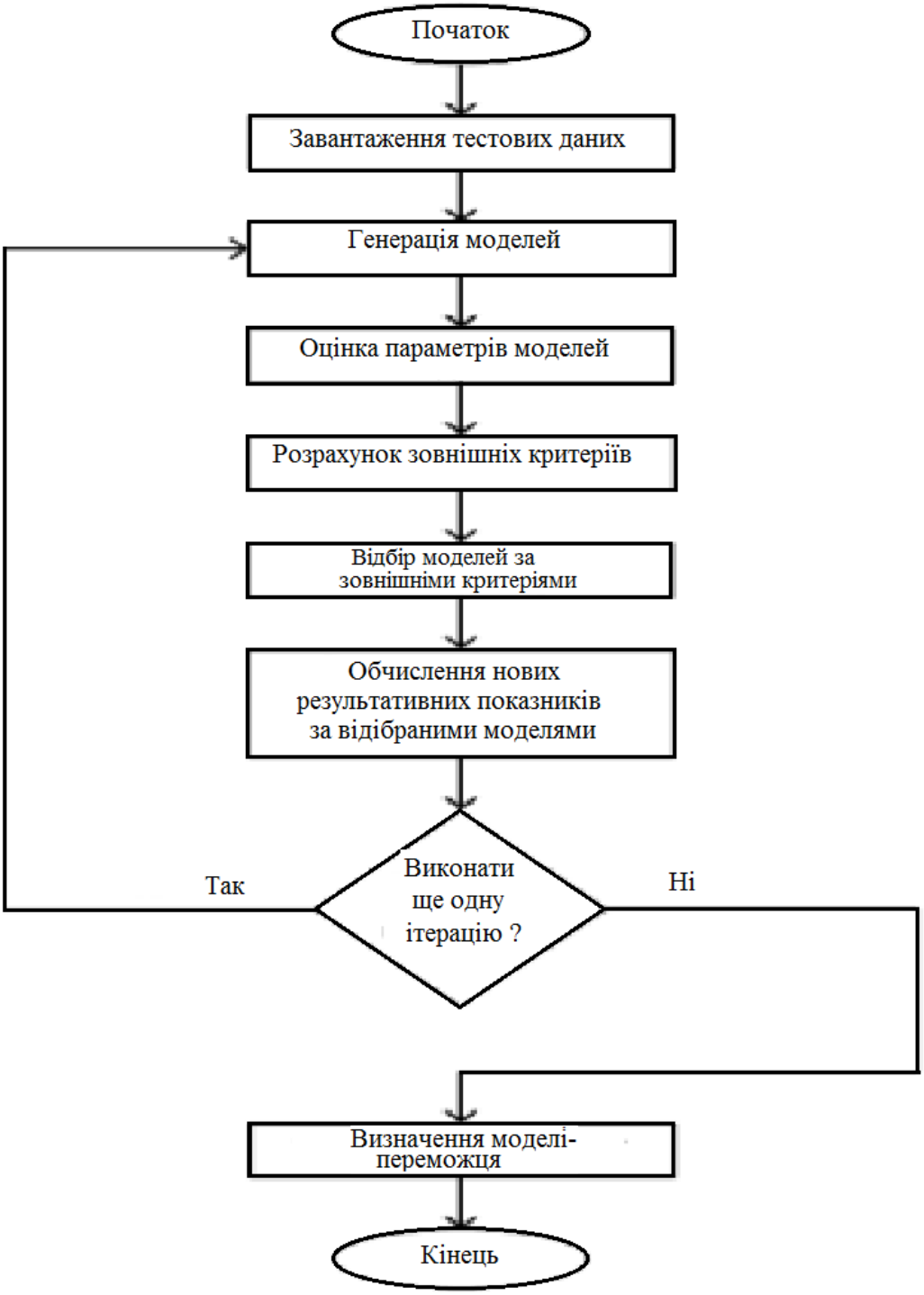

Рисунок 2.1 – Схема МГУА

2.2 Побудова та реалізація алгоритму селекції моделей

Наступним етапом після побудови залежностей є вибір критеріїв, за якими слід проводити селекцію побудованих моделей. Одним із важливих критеріїв є коефіцієнт детермінації. Тому для кожної моделі він розраховувався з допомогою модуля sklearn.liner_model, що містить метод score [12]. Даний метод

розраховує коефіцієнт детермінації та приймає на вхід масив X та масив Y.

На першій ітерації методу коефіцієнти детермінації вийшли досить високі. Код з рис. 2.2 демонструє отримані результати. Таким чином, було прийнято рішення вибирати ті моделі, у яких коефіцієнт детермінації більший або дорівнює 0.6. Таких моделей виявилося 132. Окремі з них представлені в Додатку А.

```
model: y = b0 + b1*lnX1 + b2*lnX2 + … + e
column number = (1, 2, 3, 4, 5, 6, 7)
coefficient of determination = 0.7145716325523145
```

Рисунок 2.2 –Результат обчислення коефіцієнта детермінації

На другій ітерації методу у відібрані моделі подаються результати попередньої ітерації. Для обчислення результатів використовувався модуль sklearn.liner_model, який містить метод predict [12]. Цей метод приймає на вхід матрицю X (тестові дані) і передбачає результати, використовуючи розраховані коефіцієнти побудованої лінійної моделі. Після підстановки результатів у побудовані моделі знову розраховується коефіцієнт детермінації.

Однак на другій ітерації методу виникли труднощі. Кількість передбачених результатів дорівнювала 132. Для реалізації другого етапу було необхідно знову реалізувати комбінаторний перебір порахованих значень. Але так як на першій ітерації було 9 пояснюючих факторів, на другій їх стало 132. Як наслідок, збільшення числа вхідних значень свідчить про глобальне збільшення кількості варіантів для кожної групи моделей. Наприклад, для побудови лінійної моделі, що містить 8 доданків (з урахуванням вільного члена) потрібно реалізувати перебір 132 вхідних значень без повторень. За формулою 10, яка вираховує кількість всіляких поєднань з n різних елементів по k, можна сказати, що кількість варіантів для наведеної в прикладі моделі буде дорівнює 117850651776. Такий результат занадто великий для обчислень на комп'ютері без під’єднання спеціального обладнання. Тому було прийнято рішення змінити значення критерію відбору моделей (коефіцієнта детермінації) на першій ітерації

методу.

Таким чином, були відібрані тільки ті моделі, які мали коефіцієнт детермінації більшим або рівним 0.66. Таких моделей виявилося 63. Звичайно, кількість варіантів зменшилася, але все ж таки кількість моделей, що мають 8 доданків (з урахуванням вільного члена) була досить великою (553270671 варіантів). Внаслідок цього було прийнято рішення обмежитися кількістю моделей, відібраних на другій ітерації методу. Обмеження було встановлено на 150 моделей (формула (2.3)):

$$C_n^k = \frac{n!}{(n-k)!\,k!}\,, \tag{2.3}$$

де *n* - кількість різних об'єктів, *k* - кількість об'єктів у кожному варіанті.

У формулі (2.4) наведено приклад другої ітерації МГУА:

$$\begin{aligned} Y \approx {} & b_0 + b_1(a_0 + a_1x_1 + a_2x_2 + \ldots + a_8x_8 + \varepsilon) + b_2(\overline{a_0} + \overline{a_1}lnx_1 + \overline{a_2}lnx_2 + \\ \ldots & + \overline{a_8}lnx_8 + \bar{\varepsilon}) + b_3(\widehat{a_0} + \widehat{a_1}x_3^2 + \widehat{a_2}x_7^2 + \widehat{a_3}x_8^2 + \hat{\varepsilon}) + b_4(\overline{\overline{a_0}} + \overline{\overline{a_1}}lnx_3 + \overline{\overline{a_2}}lnx_5 + \\ & + \overline{\overline{a_3}}lnx_7 + \overline{\overline{a_4}}lnx_8 + \bar{\bar{\varepsilon}}) + \ldots + b_8(\widehat{\widehat{a_0}} + \widehat{\widehat{a_1}}x_2 + \widehat{\widehat{a_2}}x_6 + \widehat{\widehat{a_3}}x_7 + \widehat{\widehat{a_4}}x_8 + \hat{\hat{\varepsilon}}) + \xi, \end{aligned} \tag{2.4}$$

де Y - результативний показник, $X_1, X_2, \ldots, X_n$ - пояснювальні фактори, $b_0, b_1, \ldots, b_n$, $a_0, a_1, \ldots, a_n$ - коефіцієнти рівнянь, ε – випадкова величина.

На другій ітерації методу у збудованих моделей знову розраховується коефіцієнт детермінації. Для порівняння на рис. 2.3 представлені старі та нові значення коефіцієнта в одній із побудованих моделей. Можна помітити, що значення збільшилося, стало ще більш наближеним до 1, що говорить про добрий результат.

```
model: y = b0 + b1x1 + b2x2 + … + e
last coefficient of determination = 0.699788493876711
new coefficient of determination = 0.7663108102825795
```

Рисунок 2.3 – Результат обчислення коефіцієнта детермінації на другій ітерації методу

Для відбору моделей другого етапу значення коефіцієнта детермінації мало перевищувати максимальне значення, котре отримане на першій ітерації методу. В результаті доводилося запам'ятовувати всі коефіцієнти детермінації у відібраних моделей першого етапу, обчислювати максимальне значення і порівнювати коефіцієнти другого етапу з цим максимумом. Максимальний коефіцієнт на першому етапі дорівнював приблизно 0.71. Тим самим, новий коефіцієнт повинен бути явно більшим за це значення. Як було сказано вище, довелося встановити обмеження на кількість моделей. Оскільки було вирішено обмежитися 150 моделями, довелося вивести формулу, яка рівномірно розраховує кількість моделей, необхідних з кожної групи.

Також, моделі перевірялися на автокореляцію залишків за допомогою бібліотеки statsmodels.stats.stattools, яка включає в себе об'єкт durbin_watson [14]. Метод приймає на вхід залишки регресійної моделі та обчислює статистику WD. Залишки було отримано за допомогою об'єкта resid.

Всі відібрані моделі також перевірялися на мультиколінеарність за допомогою бібліотеки numpy, що включає об'єкт np.linalg.cond [9]. Даний об'єкт обчислює число обумовленості матриці, що подається на вхід до цього методу.

Ще одним важливим критерієм селекції моделей є перевірка на гетероскедастичність. Перевірка здійснювалась за допомогою бібліотеки statsmodels.stats.api, що включає об'єкт het_breuschpagan [16]. Цей об'єкт перевіряє модель на наявність гетероскедастичності за допомогою тесту Бреуша-Пагана. Також на вхід подаються залишки регресійної моделі.

Схема алгоритму селекції моделей показана на рис. 2.4

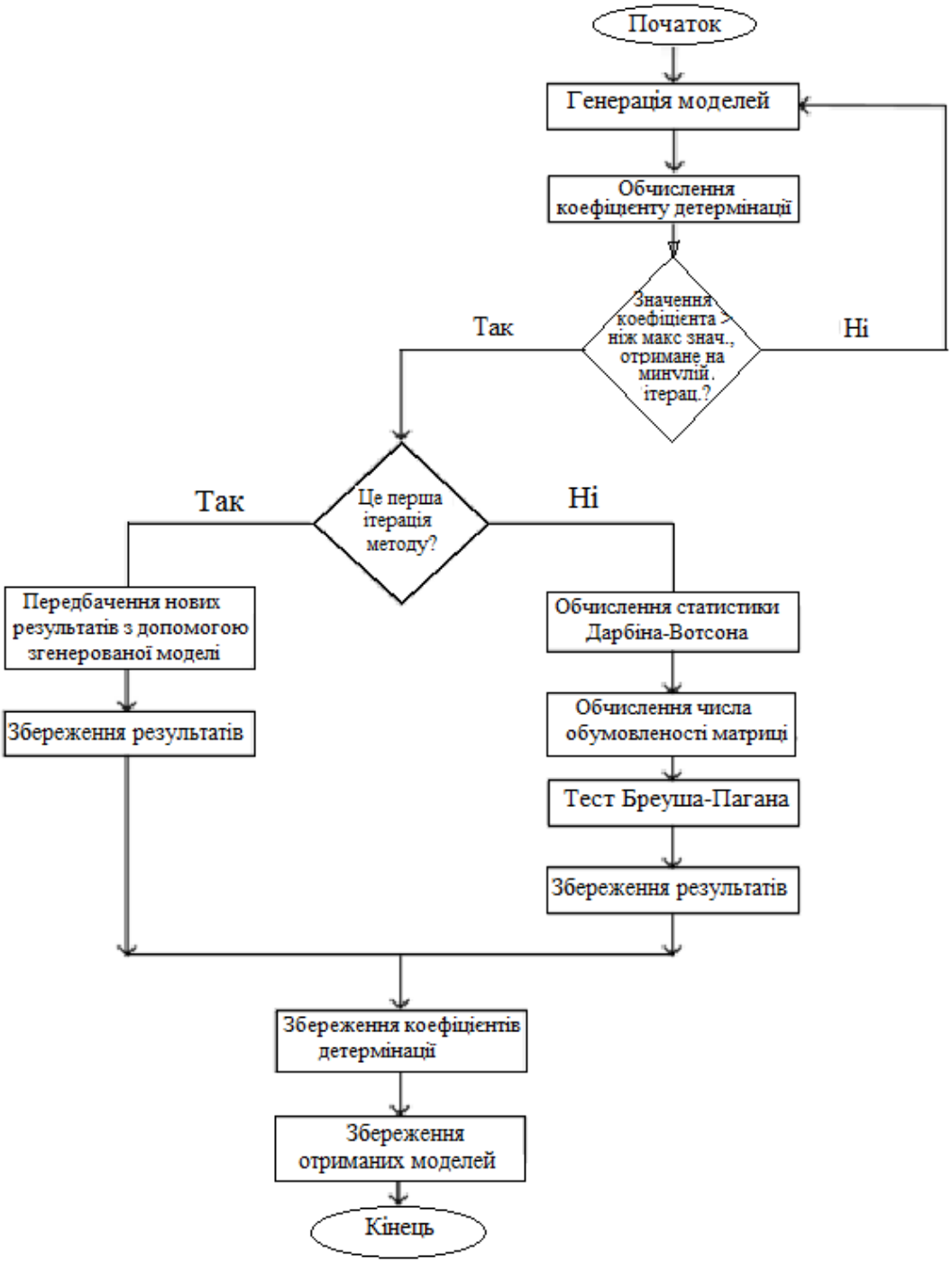

Рисунок 2.4 – Схема алгоритму селекції моделей

2.3 Результати перевірки якості збудованих моделей

Після обчислення різних перевірок було відібрано єдину модель-переможця, яка найбільш чітко описує дані. Результати перевірки кількох моделей представлені на рис. 2.5. Інші результати частково представлені у Додатку Б.

```
model: y = b0 + b1x1 + b2x2 + ... + e
column number =  (0, 1, 2, 3)
coefficient of determination =  0.7713374992777191
durbin watson =  0.784771970659844
multicollinearity =  170.52574338229897
heteroscedasticity=(6533.3380636,0.0,1852.82631282112, 0.0)

model: y = b0 + b1x1^2 + b2x2^2 + ... + e
column number =  (0, 1, 2, 3)
coefficient of determination =  0.8330648426565017
durbin watson =  1.3991873034307079
multicollinearity =  96.52344156305556
heteroscedasticity=(22436.443055907,0.0,9458.820522087175,0.0)
```

Рисунок 2.5 – Результат обчислення перевірки на автокореляцію, гетероскедастичність та мультиколінеарність

Як можна помітити, в цих моделей відсутня гетероскедастичність: значення статистики не перевищує критичне значення, що говорить про наявність гомоскедастичності. У всіх моделей, які проходили перевірку, гетероскедастичність відсутня.

Також, у моделей з рис. 2.5 є сильна мультиколінеарність (значення набагато більше 30). Такий результат цілком очевидний: у медицині багато показників сильно залежать від інших. Наприклад, з віком у людини результати вимірів дедалі частіше перевищують норму. Практично у всіх моделей, що проходили перевірку, мультиколінеарність сильно перевищує позначку 30. Тільки в одній моделі результат показав значення 30, що говорить про наявність несильної мультиколінеарності.

Щодо значень статистики DW результати обчислень кількох моделей з рис. 2.5 свідчать про наступне: у першій моделі статистика дорівнює 0.78, що говорить про зону невизначеності, тобто слід провести більше досліджень та робіт над якістю. Результат другої моделі набагато кращий: статистика наближається до 2, що говорить про відсутність автокореляції залишків. Результати інших моделей, що проходили перевірку, показали різні значення: у

більшості статистика потрапляла в зону невизначеності. Але були отримані такі значення, які наближалися до 2.

За результатами обчислень було зроблено такий висновок: для відбору моделі-переможця буде враховано статистику DW, і наскільки сильно зріс коефіцієнт детермінації після другої ітерації методу. Для реалізації наступного алгоритму потрібно запам'ятовувати статистику у кожної відібраної моделі на другому етапі методу та шукати значення найближче до 2. А також аналізувати значення коефіцієнта детермінації (шукати найбільше значення з усіх коефіцієнтів, отриманих на другій ітерації). Саме за цими двома критеріями було відібрано модель-переможця. Рис. 2.6 демонструє реалізацію описаного вище алгоритму.

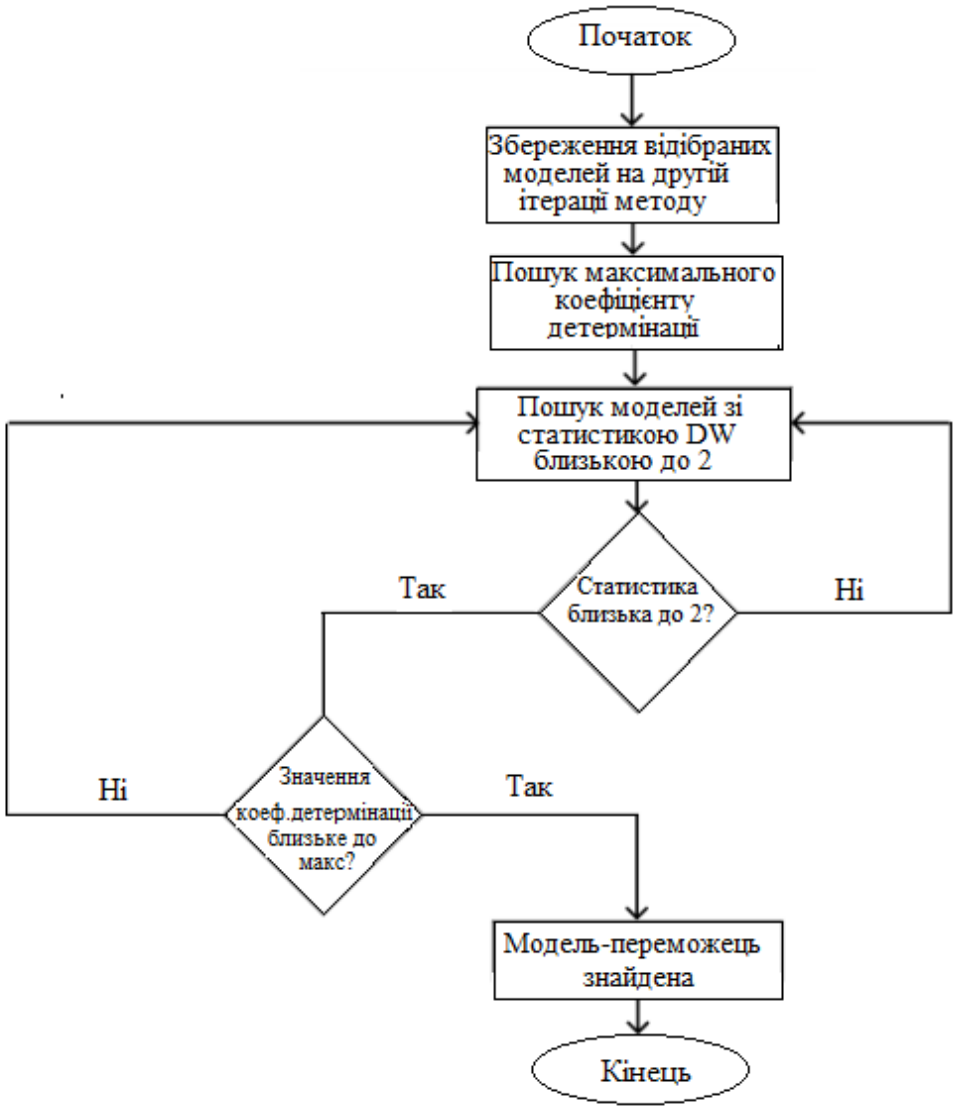

Рисунок 2.6 – Схема алгоритму визначення моделі-переможця

В результаті було обрано модель-переможець, яка наведена на рис. 2.7.

```
['model: y = b0 + b1x1^2 + b2x2^2 + ... + e' 'column number:
(0, 1, 2, 3, 9)' 'determination:' '0.8341726705594799' 'dw:'
'1.4063679738906 58''mult:''139.63290522732592' 'heteroscedasticity:'
'(22569.419198149044, 0.0, 7642.8614394938995, 0.0)']
```

Рисунок 2.7 – Результат роботи алгоритму пошуку моделі-переможця

2.4 Побудова довірчого інтервалу для прогнозованого значення

Наступним етапом виконання дослідження є прогнозування ризику розвитку ССЗ. За допомогою побудованої моделі можна передбачити рівень хвороби. Для визначення точного результату обов'язковим етапом є побудова довірчого інтервалу [4], який використовується під час інтервальної оцінки статистичних параметрів [17]. Цей інтервал із заданою ймовірністю покриває параметр, що оцінюється. Імовірність, з якою значення параметра не потрапляє у довірчий інтервал, носить назву рівня значущості α. Імовірність того, що довірчий інтервал покриває значення параметра є рівнем довіри β = 1 — α. Рис. 2.8 демонструє довірчий інтервал із рівнем значущості α.

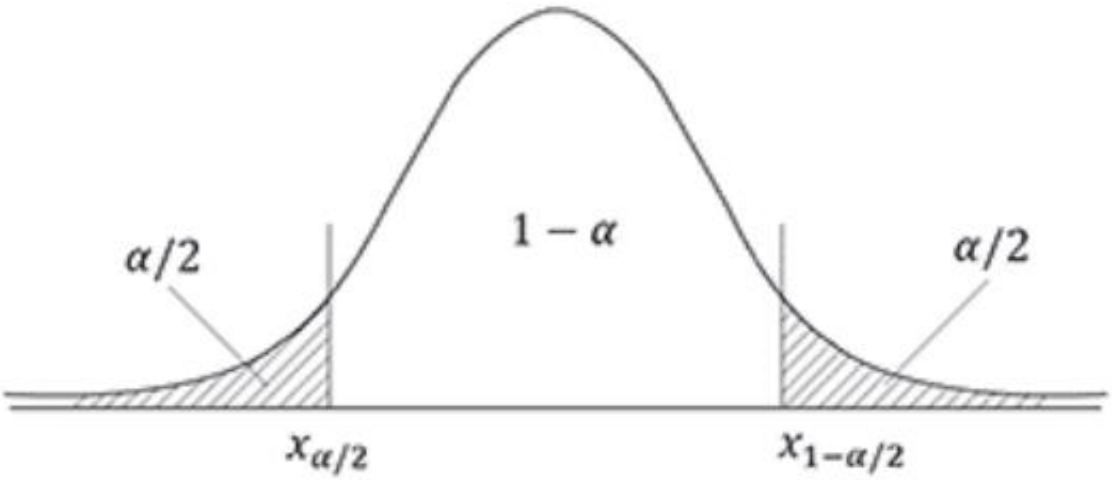

Рисунок 2.8 – Двосторонній довірчий інтервал

Зазвичай, ймовірності виходу межі довірчого інтервалу в обидві сторони рівні між собою, тому рівень значимості ділиться на 2.

Для побудови довірчого інтервалу в роботі використовувалася бібліотека statsmodel [13], яка має об'єкт summary_frame, що приймає на вхід рівень значущості a [18]. Повертає DataFrame, що містить довірчий інтервал. Було прийнято рішення взяти рівень значимості рівний 0.05, щоб довірчий інтервал мав покриття 95%.

2.5 Оформлення діалогового вікна для спілкування з користувачем

Після вибору моделі-переможця було реалізовано діалогове вікно для спілкування з користувачем та обчисленням ризику розвитку ССЗ у пацієнта. Діалогова система ставить питання, збираючи дані для обчислення результатів. Для обчислення ризику необхідно отримати такі показники:

- рівень глюкози;
- рівень холестерину;
- систолічний артеріальний тиск;
- діастолічний артеріальний тиск;
- BMI;
- статус куріння;
- підлога;
- вік;
- наявність АГ.

Рис. 2.9 демонструє збирання показників та розрахунок ризику виникнення АГ, обчислюючи при цьому довірчий інтервал для отриманого значення.

```
Ризик розвитку серцево-судинного захворювання
Введіть рівень глюкози:  5.7
Введіть рівень холестерину:  7.5
Введіть рівень систолічного тиску:  160
Введіть рівень діастолічного тиску:  80
BMI:  27
Статус паління (введіть Yes/No/Formerly):  Yes
Стать (введіть: M/F):  M
Вік:  74
Чи має пацієнт серцево-судинне захворювання? (введіть:Yes/No):
Yes
ризик розвитку серцево-судинного захворювання =  [26.70666389]
довірчий інтервал = [23.56131253627894; 29.852015249126744]
```

Рисунок 2.9 – Діалогове вікно з користувачем

Після збору показників алгоритм кодує текстові значення, замінюючи їх числовими (0 та 1). Далі створює їх масив і розраховує результативний показник Y, підставляючи зібрані дані у модель-переможець.

Для обчислення ризику потрібно запам'ятовувати параметри всіх моделей першої ітерації, що входять у модель-переможець і ці моделі. Після збору показників було реалізовано алгоритм, який розраховував, які моделі з першої ітерації входять у модель-переможець. Вибрана модель зберігається у масиві, що містить номери стовпців матриці, котра містить результати першого етапу алгоритму. Таким чином алгоритм перебирає ці стовпці, звертаючись до масиву з порахованими коефіцієнтами моделей першого етапу. Номери стовпців матриці з результатами першого етапу відповідали порядковому номеру матриці, що містить коефіцієнти моделей першої ітерації. За допомогою цього вдалося підставити зібрані у користувача значення у всі моделі, що входять у модель-переможець.

Тим самим, отримавши нові результати моделей першої ітерації, обчислювалося підсумкове значення, яке є прогнозованим ризиком ССЗ. Нові підраховані результати підставлялися в модель-переможець і розраховувалося кінцеве значення, яке було відповіддю.

Алгоритм обчислення представлений на рис. 2.10.

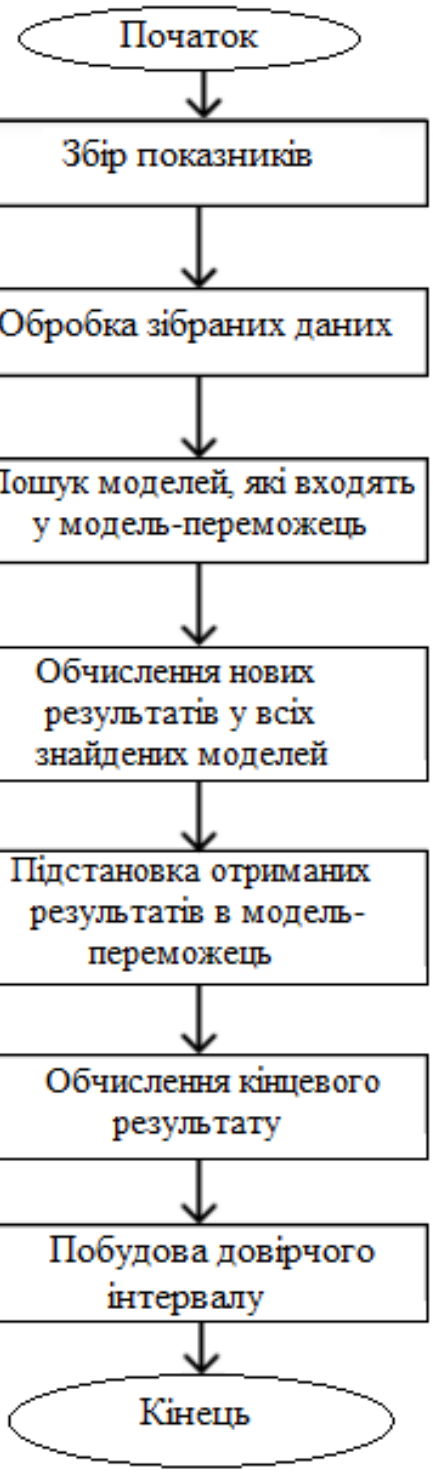

Рисунок 2.10 – Алгоритм розрахунку ризику розвитку АГ

РОЗДІЛ 3. ТЕХНІЧНА РЕАЛІЗАЦІЯ СИСТЕМИ

3.1 Архітектура системи

Система складається із двох додатків. Перша програма розроблена на Java, а друга програма реалізована на Python. Java програма відповідає за спілкування з користувачем, де виконуються процеси аутентифікації та генерації HTML - сторінок. Python-частина відповідає за зберігання алгоритму МН та моделі. Для управління БД застосовується реляційна СУБД PostgreSQL. Схема взаємодії користувача із системою представлена на рис. 3.1.

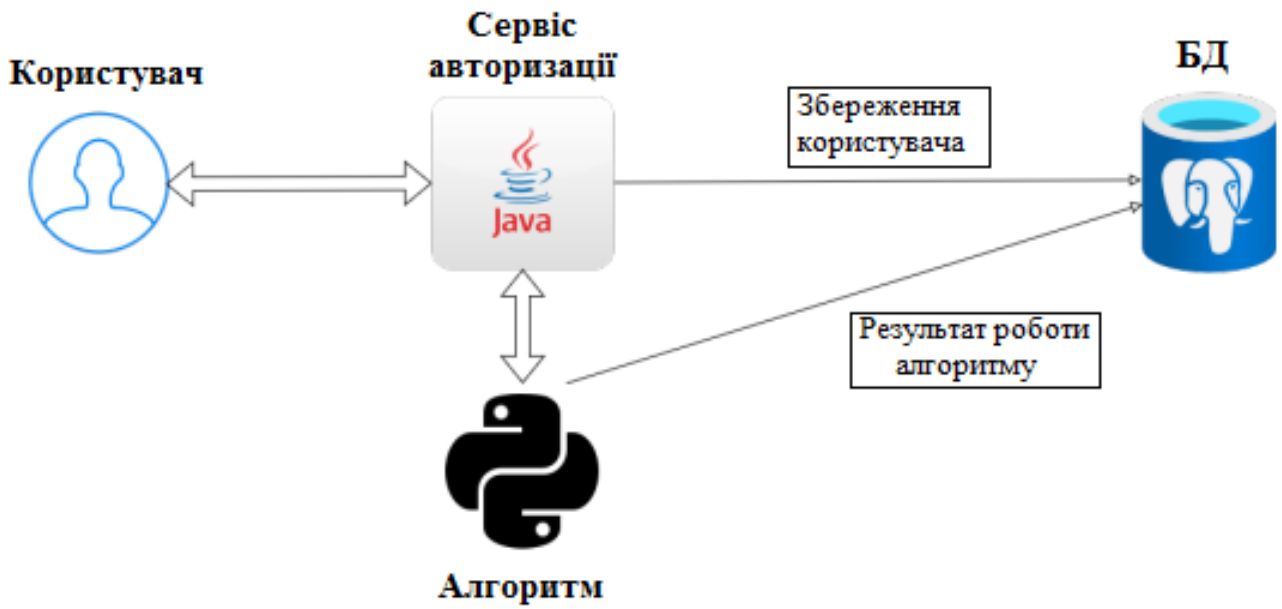

Рисунок 3.1 – Схема взаємодії API

Взаємодія з клієнтом здійснюється відповідно до принципів REST - архітектури. Обмін інформацією між Java програмою та Python програмою здійснюється у форматі JSON, де запити та відповіді представлені у вигляді JSON- об'єктів.

3.2 Реалізація Java частини

Java програма реалізована з використанням фреймворку Spring Boot [19],

який є популярним інструментом для створення серверних додатків, написаних на Java.

Відповідальність за безпеку системи доручається Spring Security. Цей фреймворк надає різні функції та інструменти для забезпечення безпеки програми [20]. У цьому системі доступом до функціональності здійснюється через токен.

Для доступу до системи користувач повинен надати дійсний токен. Після отримання токена та перевірки його дійсності, Spring Security виконує аутентифікацію користувача та встановлює відповідний аутентифікаційний контекст. Потім програма може використовувати цей контекст для перевірки дозволів та забезпечення безпечної взаємодії із системою.

Використання токенів для доступу до системи забезпечує зручність та безпеку. Токени можуть бути видані після успішної автентифікації користувача та мають обмежений термін дії. Це дозволяє контролювати доступ та запобігати несанкціонованому використанню системи.

Таким чином, завдяки Spring Security та механізму токенів, система забезпечує безпеку та контроль доступу, забезпечуючи лише авторизованим користувачам можливість взаємодії з додатком.

Для надсилання та отримання відповідей від програми Python використовується бібліотека RestTemplate. Бібліотека ModelMapper застосовується для перетворення та зіставлення даних між сутностями та JSON-об'єктами. Додаток має валідацію форм за допомогою Spring Boot Validation.

Результатом роботи програми є HTML -сторінка, яка генерується за допомогою шаблонізатора Freemarker, який надає зручні засоби для створення динамічних шаблонів веб-сторінок. Він дозволяє розділяти логіку програми від представлення, що сприяє підвищенню читання та підтримуваності коду, також він може інтегруватися зі Spring. Під час розробки візуальної частини використовувалися мови програмування JavaScript, що дозволяє додавати інтерактивність на веб-сторінці, та CSS, який слугує для визначення зовнішнього вигляду та стилю веб-сторінок. Для створення динамічних графіків та діаграм використовувалася JavaScript -бібліотека Chartjs [21]. Вона надає зручний спосіб

відображення даних у вигляді графіків, що корисно для відстеження трендів, аналізу даних та подання результатів.

У функціональні можливості програми входить:

– реєстрація користувачів (рис. 3.2);

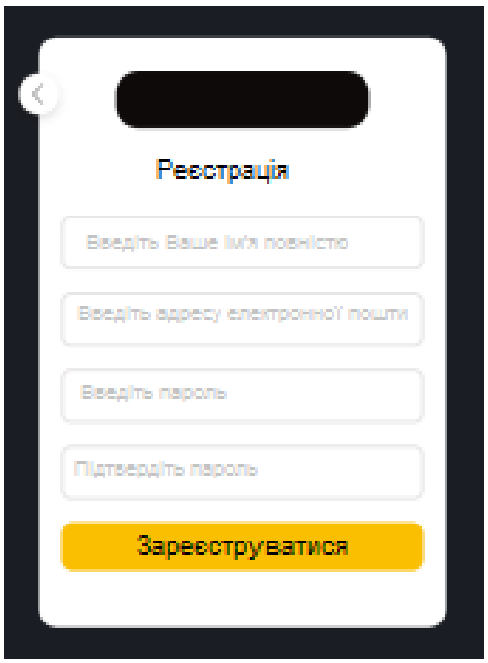

Рисунок 3.2 – Форма реєстрації користувача

– автентифікація користувачів за логіном та паролем (рис. 3.3);

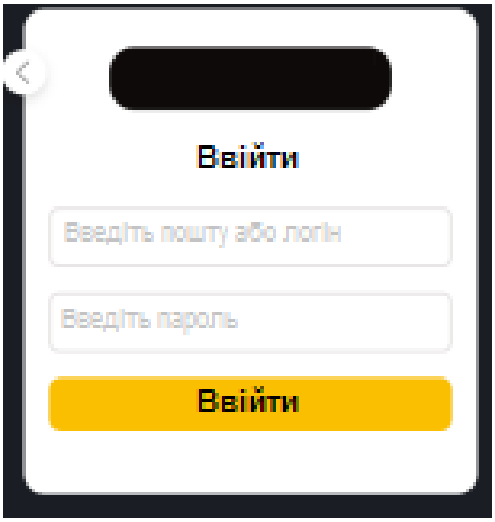

Рисунок 3.3 – Форма автентифікації користувача

– навігація по сайту: користувач може переміщатися по розділах за допомогою меню (рис. 3.4);

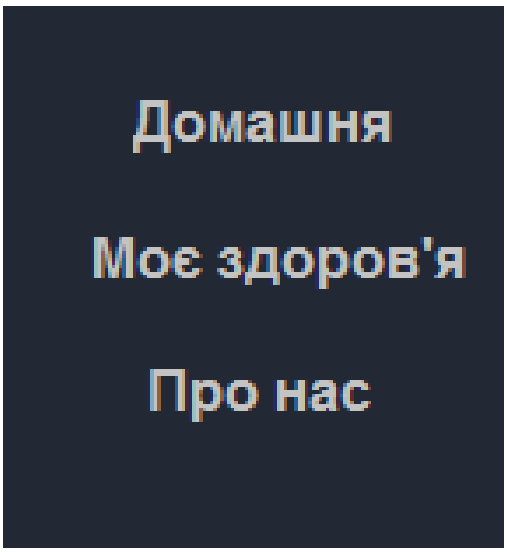

Рисунок 3.4 – Розкрите меню

– прогноз рівня ризику виникнення АГ: користувач може отримати результат прогнозу моделі, заповнивши форму зі своїми показниками здоров'я.

3.2 Реалізація Python частини

Python програми розроблений із застосуванням фреймворку FastAPI [22], відомий своєю здатністю забезпечувати високу продуктивність та ефективність в обробці запитів. Основний endpoint /predict/user/{user_id} надає можливість авторизованим користувачам отримати результат роботи моделі через програму Java , тому що Python додаток є внутрішнім і закритий для зовнішнього доступу. Було використано бібліотеку SQLAlchemy [23], яка є потужним інструментом для роботи з БД у мові програмування Python. Вона надає абстракцію над різними СУБД. Бібліотека Requests є однією з найпопулярніших бібліотек для роботи з HTTP -запитами у мові програмування Python. Вона надає інтерфейс для надсилання запитів на віддалені сервери та отримання відповідей.

3.4 Складання та розгортання програми

За складання Java програми відповідає Apache Maven - це інструмент управління проектами мовою Java. Він надає засоби автоматизації складання проекту, управління залежностями, розгортання та підтримує використання плагінів, які розширюють його функціональність. У Python частині використовується Uvicorn, розроблений для запуску та обробки веб-додатків.

У системі процес розгортання застосунків на сервері здійснюється на віртуальній машині, використовуючи Docker, на базі операційної системи, подібної до UNIX. При розгортанні створюються два контейнери, що містять Java та Python програми, а також контейнер із СУБД PostgreSQL.

Розглянемо основні команди для запуску всіх програм:

– перед запуском контейнерів потрібно виконати складання Java програми в папці Authorization $ mvn clean package;

– далі запускаємо контейнери за допомогою файлу docker-compose.yml у корені проекту $ docker-compose up --build –d;

– перевіряємо статус контейнерів $ docker ps;

Якщо всі перераховані вище кроки були виконані і статуси всіх контейнерів дорівнює UP, то вже можна почати взаємодію з сайтом, ввівши посилання http://localhost:8080 у пошуковому рядку.

РОЗДІЛ 4. БЕЗПЕКА ЖИТТЄДІЯЛЬНОСТІ, ОСНОВИ ОХОРОНИ ПРАЦІ

4.1 Класифікація шкідливих та небезпечних виробничих факторів

Шкідливий виробничий фактор – небажане явище, яке супроводжує виробничий процес і вплив якого на працюючого може призвести до погіршення самопочуття, зниження працездатності, захворювання, виробничо зумовленого чи професійного, і навіть смерті, як результату захворювання. Небезпечний виробничий фактор – небажане явище, яке супроводжує виробничий процес і дія якого за певних умов може призвести до травми або іншого раптового погіршення здоров'я працівника (гострого отруєння, гострого захворювання) і навіть до раптової смерті [23].

Поділ несприятливих чинників виробничого середовища на шкідливі та небезпечні зумовлене різним характером їх дії на людський організм, тим, що вони потребують різних заходів та засобів для боротьби з ними та профілактики викликаних ними ушкоджень, а також рядом причин організаційного характеру. В той же час між шкідливими та небезпечними виробничими факторами інколи важко провести чітку межу. Один і той же чинник може викликати травму і профзахворювання (наприклад, високий рівень іонізуючого або теплового випромінювання може викликати опік або навіть призвести до миттєвої смерті, а довготривала дія порівняно невисокого рівня цих же факторів – до хвороби; пилинка, що потрапила в око, спричиняє травму, а пил, що осідає в легенях, – захворювання, що зветься пневмоконіоз). Через це всі несприятливі виробничі чинники часто розглядаються як єдине поняття – небезпечний та шкідливий виробничий фактор (НШВФ) [24]. За своїм походженням та природою дії всі НШВФ можна поділити на 5 груп: фізичні, хімічні, біологічні, психофізіологічні та соціальні До фізичних НШВФ відносяться машини та механізми або їх елементи, а також вироби, матеріали, заготовки тощо, які рухаються або обертаються; конструкції, які руйнуються; системи, устаткування або елементи обладнання, які знаходяться під підвищеним тиском; підвищена запиленість та загазованість повітря; підвищена або понижена температура повітря, поверхонь

приміщення, обладнання, матеріалів; підвищені рівні шуму, вібрації, ультразвуку, інфразвуку; підвищений або понижений барометричний тиск та його різкі коливання; підвищена та понижена вологість; підвищена швидкість руху та підвищена іонізація повітря; підвищений рівень іонізуючих випромінювань; підвищене значення напруги в електричній мережі; підвищені рівні статичної електрики, електромагнітних випромінювань; підвищена напруженість електричного, магнітного полів; відсутність або нестача світла; недостатня освітленість робочої зони; підвищена яскравість світла; понижена контрастність; прямий та віддзеркалений блиск; підвищена пульсація світлового потоку; підвищені рівні ультрафіолетової та інфрачервоної радіації; гострі крайки, зачипки, шершавість на поверхні заготовок, інструментів та обладнання; розташування робочого місця на значній висоті відносно землі (підлоги); слизька підлога; невагомість.

Хімічні НШВФ:

- за характером дії на організм людини поділяються на токсичні, задушливі, наркотичні, подразнюючі, сенсибілізуючі, канцерогенні, мутагенні та такі, що впливають на репродуктивну функцію;

- за шляхами проникнення в організм людини поділяються на такі, що потрапляють через: 1) органи дихання; 2) шлунково-кишковий тракт; 3) шкіряні покриви та слизова оболонка;

- які перебувають у різному агрегатному стані: 1)твердому 2)газоподібному 3)рідкому.

Біологічні НШВФ – це: - патогенні мікроорганізми (бактерії, віруси, рикетсії, спірохети, грибки, найпростіші) та продукти їхньої життєдіяльності; - макроорганізми (тварини та рослини) та продукти їхньої життєдіяльності. До психофізіологічних НШВФ відносяться фізичні (статичні та динамічні) перевантаження і нервово-психічні перевантаження (розумове перенапруження, перенапруження аналізаторів, монотонність праці, емоційні перевантаження). 6 Соціальні НШВФ – це неякісна організація роботи, понаднормова робота, змушеність праці в колективі з поганими відносинами між його членами, соціальна ізольованість з відривом від сім_ї, зміна біоритмів, незадоволеність

роботою, фізична та/або словесна образа та її ризик, насильство та його ризик. Один і той же НШВФ за природою своєї дії може належати водночас до різних груп.

4.2 Вплив вібрації на людину

Вібрація - це механічні коливання пружних тіл або коливальні рухи механічних систем. Для людини вібрація є видом механічного впливу, який має негативні наслідки для організму [24].

Причиною появи вібрації є неврівноважені сили та ударні процеси в діючих механізмах. Створення високопродуктивних потужних машин і швидкісних транспортних засобів при одночасному зниженні їх матеріалоємності неминуче призводить до збільшення інтенсивності і розширення спектру вібраційних та віброакустичних полів. Цьому сприяє також широке використання в промисловості і будівництві високоефективних механізмів вібраційної та віброударної дії.

Дія вібрації може приводити до трансформування внутрішньої структури і поверхневих шарів матеріалів, зміни умов тертя і зносу на контактних поверхнях деталей машин, нагрівання конструкцій. Через вібрацію збільшуються динамічні навантаження в елементах конструкцій, стиках і сполученнях, знижується несуча здатність деталей, ініціюються тріщини, виникає руйнування обладнання. Усе це приводить до зниження строку служби устаткування, зростання імовірності аварійних ситуацій і зростання економічних витрат. Вважають, що 80% аварій в машинах і механізмах здійснюється внаслідок вібрації. Крім того, коливання конструкцій часто є джерелом небажаного шуму. Захист від вібрації є складною і багатоплановою в науково-технічному та важливою у соціальноекономічному відношеннях проблемою нашого суспільства [25].

Вплив вібрації на людину залежить від її спектрального складу, напрямку дії, прикладення, тривалості впливу, а також від індивідуальних особливостей людини. При оцінці вібраційного впливу потрібно враховувати, що коливальні процеси притаманні живому організму. В основі серцевої діяльності і кровообігу

та біострумів мозку лежать ритмічні коливання. Внутрішні органи людини можна розглядати як коливальні системи з пружними зв'язками. Частоти їх власних коливань лежать у діапазоні 3..6 Гц. Частоти власних коливань плечового пояса, стегон і голови щодо опорної поверхні (положення стоячи) складають 4...6 Гц, голови щодо плечей (положення сидячи) 25...30 Гц.

При впливі на людину зовнішніх коливань (хитавиці, струсів, вібрації) відбувається їхня взаємодія з внутрішніми хвильовими процесами, виникнення резонансних явищ. Так, зовнішні коливання частотою менш 0,7 Гц утворюють хитавицю і порушують у людини нормальну діяльність вестибулярного апарата. Інфразвукові коливання (менш 16 Гц), впливаючи на людину, пригнічують центральну нервову систему, викликаючи почуття тривоги, страху. При певній інтенсивності на частоті 6..7 Гц інфразвукові коливання, втягуючи у резонанс внутрішні органи і систему кровообігу, здатні викликати травми, розриви артерій, тощо [20].

Вібрація, що діє на людину, має широкий діапазон – від десятих часток одного до декількох тисяч Гц. Характерними ознаками шкідливого впливу вібрації на людину є можливі зміни у функціональному стані: підвищена втома, збільшення часу моторної реакції, порушення вестибулярної реакції. Медичними дослідженнями встановлено, що вібрація є подразником периферичних нервових закінчень, розташованих на ділянках тіла людини, що сприймають зовнішні коливання. Адекватним фізичним критерієм оцінки її впливу на організм людини є коливальна енергія, що виникає на поверхні контакту, а також енергія, поглинена тканинами і передана опорно-руховому апарату та іншим органам. У результаті впливу вібрації виникають нервовосудинні розлади, ураження кістково-суглобної та інших систем організму. Відзначаються, наприклад, зміни функції щитовидної залози, сечостатевої системи, шлунково-кишкового тракту. Так, медичні дослідження показали , що у працюючих в умовах вібрації відбуваються значні зміни кістковосуглобної системи, які виражаються у функціональній перебудові кісткової тканини, регіональному остеопорозі, кістовидних утвореннях у кістках, асептичному некрозі кісток, хронічних

переломах. Відзначається, що терміни виникнення змін у кістках у працівників вібраційних професій коливається в межах від 6-8 місяців до 2-5 років.

Шкідливість вібрації збільшується при одночасному впливі на людину таких факторів, як знижена температура, підвищений шум, запиленість повітря, тривала статична напруга тощо. Сучасна медицина розглядає виробничу вібрацію як могутній стрес-фактор, що має негативний вплив на психомоторну працездатність, емоційну сферу і розумову діяльність, підвищує ймовірність виникнення різних захворювань і нещасних випадків. Особливо небезпечний тривалий вплив вібрації для жіночого організму. Широкий комплекс патологічних відхилень, викликаний впливом вібрації на організм людини, кваліфікується як віброзахворювання [25].

Вібрація як фізичний чинник виробничого середовища спостерігається в металообробній, гірничодобувній, металобудівній, машинобудівній, авіаційній та інших галузях народного господарства. Джерелом вібрації можуть бути різні механізми, вібраційне устаткування, віброінструменти, акустичні системи, транспортні та сільськогосподарські машини.

Загальна вібрація поділяється на транспортну вібрацію, яка діє на людину на робочих місцях в транспортних засобах (трактори сільськогосподарські та промислові, самохідні сільськогосподарські машини (комбайни), тягачі, грейдери ті інші); транспортно-технологічну вібрацію, яка діє на людину на робочих місцях машин з обмеженою рухливістю (екскаватори, крани промислові та будівельні, гірничі комбайни, транспорт виробничих приміщень та інші) та технологічну вібрацію, яка діє на людину на робочих місцях стаціонарних машин чи передається на робочі, де немає джерел вібрації (верстати та метало-деревообробне, пресувально-ковальське обладнання, ливарні машини, електричні машини, насосні агрегати та вентилятори, обладнання для буріння свердловин, бурові верстати, машини для тваринництва, очищення та сортування зерна (у тому числі сушарні), обладнання промисловості будматеріалів (крім бетоноукладачів), установки хімічної та нафтохімічної промисловості та інші.

Оператори машин, які зазнають у процесі трудової діяльності впливу вібрації, підлягають попереднім та періодичним медичним оглядам відповідно

до Порядку проведення медичних оглядів працівників певних категорій, затвердженого Наказом МОЗ України від 21.05.2007 р. №246. Обов'язкові попередні (під час прийняття на роботу) та періодичні (протягом трудової діяльності) медичні огляди дозволять визначити стан здоров'я працівника та можливість виконання без погіршення стану здоров'я професійних обов'язків, своєчасно виявити ранні ознаки хронічного професійного захворювання, забезпечує динамічне спостереження за станом здоров`я в умовах дії шкідливих та небезпечних факторів і трудового процесу, вирішує питання щодо можливості продовжувати роботу в умовах дії шкідливих та небезпечних факторів і трудового процесу [25].

За результатами періодичних медичних оглядів роботодавець забезпечує проведення відповідних оздоровчих заходів Заключного акта у повному обсязі та усуває причини, що призводять до професійних захворювань. Організовує проведення лабораторних досліджень умов праці на робочих місцях та вживає заходів до усунення небезпечних і шкідливих для здоров'я виробничих факторів.

До роботи операторами машин допускаються особи не молодші 18 років, які пройшли попередній медичний огляд, мають відповідну кваліфікацію та ознайомлені з характером впливу вібрації на організм.

ВИСНОВКИ

В результаті виконання роботи було реалізовано ІС, побудовано модель, що описує дані медичного характеру, реалізовано інтерфейс для спілкування з користувачем та збирання показників для передбачення ризику розвитку АГ, а також алгоритм прогнозування ризику розвитку захворювання.

Було виконано такі завдання:

– проведено попередню обробку даних;
– підібрано сімейство для побудови моделей;
– вибрано необхідні критерії для відбору моделей;
– реалізований інструмент для побудови моделей;
– реалізовано інструмент для перевірки якості побудованих моделей;
– обрано єдину модель-переможця;
– реалізовано людино-машинний інтерфейс для спілкування з користувачем;
– побудований довірчий інтервал для прогнозованого значення.

У перспективі планується провести кілька додаткових ітерацій МГУА для покращення якості прогнозування ризику розвитку захворювання, провести дослідження щодо виявлення відмінностей у станах здоров'я між міським та сільським населенням, удосконалити людино-машинний інтерфейс для більш комфортного спілкування з користувачем.

Також планується провести дослідження щодо усунення мультиколінеарності та автокореляції та реалізувати ще кілька перевірок моделей.

ПЕРЕЛІК ДЖЕРЕЛ

1. Оцінка серцево-судинного ризику. [Електронний ресурс]. – Режим доступу: https://empendium.com/ua/manual/chapter/B72.I.D.3. (Дата звертання: 28.03.2024)
2. Методи моделювання складних систем і процесів: Навчальний посібник. [Електронний ресурс]. – Режим доступу: https://ela.kpi.ua/bitstream/123456789/50988/1/Metody_modeliuvannia.pdf (Дата звертання: 28.03.2024)
3. Пасічник В.В., Виклюк Я.І., Камінський Р.М. Моделювання складних систем. Посібник. Львів: Видавництво "Новий Світ - 2000". 2017. 404 с.
4. Green W.H. Econometric analysis – 8-th Edition, Pearson. 2017. – 1176 p.
5. Економетрика ^підручник / За ред. О. І. Черняка. – Миколаїв : МНАУ, 2015. – 414 с.
6. Літнарович Р.М.Побудова і дослідження математичної моделі за джерелами експериментальних даних методами регресійного аналізу. Навчальний посібник, МЕГУ, Рівне, 2011.-140 с.
7. Прикладна економетрика : навч. посіб. : у двох частинах. / Л. С. Гур'янова, Т. С. Клебанова, С. В. Прокопович та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2016. – 235 с.
8. Волошин О.Р., Галайко Н.В. Економетрія. навч. посібник / О. Волошин, Н. Галайко. – Львів: Львівський державний університет внутрішніх справ, 2012. – 192 с.
9. Документація Numpy. [Електронний ресурс] - Режим доступу: https://numpy.org/doc/stable/reference/ (дата звертання: 08.05.2024)
10. Бібліотека Pandas. Посібник із використання pandas для аналізу великих наборів даних. [Електронний ресурс] - Режим доступу: https://habr.com/ru/company/ruvds/blog/442516/ (дата звертання: 22.04.2024)
11. Бібліотека Itertools. [Електронний ресурс] - Режим доступу: https://habr.com/ru/company/otus/blog/529356/ (дата звертання: 07.05.2024)
12. Документація Scikit-learn. sklearn.liner_model. LinearRegression [Електронний ресурс] - Режим доступу: https://scikit-

learn.org/stable/modules/generated/sklearn.linear_model.LinearRegr ession.html (дата звертання: 30.04.2024)

13. Документація Statsmodels. Introduction — statsmodels. [Електронний ресурс] - Режим доступу: https://www.statsmodels.org/stable/index.html (дата звертання: 02.05.2024)

14. Аналіз даних з Python [Електронний ресурс] - Режим доступу: https://www.freecodecamp.org/ukrainian/learn/data-analysis-with-python/ (дата звертання: 20.04.2024)

15. Регресійний аналіз за допомогою Python. [Електронний ресурс] - Режим доступу: https://www.researchgate.net/publication/365607989_ Regresijnij_analiz_za_dopomogou_Python (дата звертання: 20.04.2024)

16. Регресійна діагностика [Електронний ресурс] - Режим доступу: https://www.statsmodels.org/devel/examples/notebooks/generated/regression_dia gnostics.html (дата звертання: 14.05.2024)

17. Гур'янова Л.С., Клебанова Т.С., Сергієнко О.А., Прокопович С.В. Економетрика. Навчальний посібник -Харків: Вид. ХНЕУ ім. С. Кузнеця, 2015. – 389 с.

18. Довірчі інтервали до лінійного регресійного аналізу [Електронний ресурс] - Режим доступу: http://ni.biz.ua/9/9_15/9_155404_doveritelnie-intervali-v-lineynom-regressionnom-analize.html (дата звертання: 24.05.2024)

19. Spring Boot [Електронний ресурс]. – Режим доступу: https://spring.io/projects/springboot (дата звертання: 23.05.2024)

20. Spring Security [Електронний ресурс]. – Режим доступу: https://spring.io/projects/spring-security (дата звертання: 23.05.2024)

21. Chart.js [Електронний ресурс]. – Режим доступу: https://www.chartjs.org/ (дата звертання: 23.05.2024)

22. FastAPI [Електронний ресурс]. – Режим доступу: https://fastapi.tiangolo.com/ (дата звертання: 23.05.2024)

23. SQLAlchemy 2.0 Documentation [Електронний ресурс]. – Режим доступу: https://docs.sqlalchemy.org/en/20/ (дата звертання: 23.05.2024)

24. Зеркалов Д.В. Безпека життєдіяльності та основи охорони праці.

Навчальний посібник. К.: «Основа». 2016. – 267 с.

25. Яремко З. М. Безпека життєдіяльності: Навч. посіб. — Львів., 2005. – 301 с.

26. Желібо Є. П. Заверуха Н.М., Зацарний В.В. Безпека життєдіяльності. Навчальний посібник. – К.: Каравела, 2004. -328 с.

ДОДАТКИ

Додаток А

Фрагмент лістингу розробленого програмного забезпечення алгоритму селекції моделей

```
model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (1,  2, 6,     7)
coefficient of determination =  0.6841932519937183

model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (2,  5, 6,     7)
coefficient of determination =  0.6630658109333778

model: y = b0 + b1lnx1 + ... + e
column number = (1, 2, 6, 7)
coefficient of determination =  0.6929842261170165

model: y = b0 + b1lnx1 + ... + e
column number = (2, 5, 6, 7)
coefficient of determination =  0.6802458419820714

model: y = b0 + b1x1 A 2 + b2x2 A 2 + ... + e
column number =   (1, 2, 6, 7)
coefficient of determination =  0.663527253603688

model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (0,  1, 2, 6,  7)
coefficient of determination =  0.6842017716769406

model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (0,  2, 5, 6,  7)
coefficient of determination =  0.6642123415641771

model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (1,  2, 3, 6,  7)
coefficient of determination =  0.6842114525503712

model: y = b0     +    b1x1      + b2x2 +  ...  +    e
column number     =    (1,  2, 4, 6,  7)
coefficient of determination =  0.6841932668429347

model: y = b0     +    blxl      + b2x2 +  ...  +    e
column number     =    (2,  3, 5, 6,  7)
coefficient of determination =  0.6630679731344263
```

```
model: y = b0     +    blxl      + b2x2 +  ...  +    e
column number     =    (2,  4, 5, 6,  7)
coefficient of determination =  0.663270047019366

model: y = b0 + bllnxl + ... + e
column number = (0, 1, 2, 6,7)
coefficient of determination =  0.6929842287374794

model: y = b0 + bllnxl + ... + e
column number = (0, 2, 5, 6, 7)
coefficient of determination =  0.6815520325612212

model: y = b0 + bllnxl + ... + e
column number = (1, 2, 3, 6, 7)
coefficient of determination =  0.6929877463926399

model: y = b0 + bllnxl + ... + e
column number = (l, 2, 4, 6, 7)
coefficient of determination =  0.6930026364564911

model: y = b0 + bllnxl + ... + e
column number = (l, 2, 5, 6, 7)
coefficient of determination =  0.7145458487680292

model: y = b0 + bllnxl + ... + e
column number = (2, 3, 5, 6, 7)
coefficient of determination =  0.6802460641057484

model: y = b0 + bllnxl + ... + e
column number = (2, 4, 5, 6, 7)
coefficient of determination =  0.6803676559954344

model: y = b0 + blxl A 2 + b2x2 A 2 + ... + e
column number =   (0, l, 2, 6, 7)
coefficient of determination =  0.6635516614348523

model: y = b0 + blxl A 2 + b2x2 A 2 + ... + e
column number =   (l, 2, 3, 6, 7
coefficient of determination =  0.65346316672312245

…….
```

Додаток Б

Фрагмент лістингу розробленого програмного забезпечення перевірки якості побудованих моделей

```
  model: y = b0 + b1x1 + b2x2 + ... + e
  column number = (0, 1, 2, 3)
  coefficient of determination =  0.7713374992777191
  durbin watson =  0.784771970659844
  multicollinearity =  170.52574338229897
  heteroscedasticity =  (6533.338063686955, 0.0,
1852.8263128211297, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =  (0, 1, 2, 17)
  coefficient of determination =  0.7713058283467903
  durbin watson =  0.7846431420254355
  multicollinearity =  170.50736062118662
  heteroscedasticity =  (6533.478941800608, 0.0,
1852.8716375748006, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number = (0, 1, 3, 4)
  coefficient of determination =  0.7704753264751367
  durbin watson =  0.7181044345481289
  multicollinearity =  132.532461913866
  heteroscedasticity =  (5907.237658272569, 0.0,
1653.9541101846846, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =  (0, 1, 3, 12)
  coefficient of determination =  0.7713373832209358
  durbin watson =  0.7847653445798339
  multicollinearity =  170.52915540134364
  heteroscedasticity =  (6533.529212847071, 0.0,
1852.8878113558274, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =  (0, 1, 3, 14)
  coefficient of determination =  0.7712737651513251
  durbin watson =  0.7840064363701589
  multicollinearity =  170.2635972916152
  heteroscedasticity =  (6525.987567206162, 0.0,
1850.4618001649435, 0.0)
  durbin watson =  0.7809957384792033
  multicollinearity =  169.81152087023295
```

```
  heteroscedasticity =   (6483.395408690576, 0.0,
1836.774794644436, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =   (0, 1, 3, 19)
  coefficient of determination =  0.7703874657382406
  durbin watson =   0.7173757129760095
  multicollinearity =    131.88992134332045
  heteroscedasticity =   (5890.399613219893, 0.0,
1648.675584550774, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e column number =    (0, 1,
3, 20)
  coefficient of determination =  0.7705620469026461
  durbin watson =   0.7187293842295923
  multicollinearity =    131.4029996301918
  heteroscedasticity =   (5877.034461553646, 0.0,
1644.4883392085576, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =   (0, 1, 3, 21)
  coefficient of determination =  0.7703719862505629
  durbin watson =   0.7168728710297866
  multicollinearity =    132.21642619941142
  heteroscedasticity =   (5882.757203884362, 0.0,
1646.2809718306667, 0.0)

  model: y = b0 + b1x1 + b2x2 + ... + e
  column number =   (0, 1, 3, 32)
  coefficient of determination =  0.7712736518937162
  durbin watson =   0.7839994305375785
  multicollinearity =    170.26715740718905
  heteroscedasticity =   (6526.1879636022995, 0.0,
1850.5262543275621, 0.0)
heteroscedasticity =    (22436.44305590731, 0.0,
9458.820522087175, 0.0)
  model: y = b0 + b1x1 A 2 + b2x2 A 2 + ... + e column number =
     (0, 1, 2, 4)
  coefficient of determination =  0.7761694826952836
  durbin watson =   1.2174866015693233
  multicollinearity =    99.98293329094321
  heteroscedasticity =   (8200.581152630473, 0.0,
2408.2874923806953, 0.0)
  model: y = b0 + b1x1 A 2 + b2x2 A 2 + ... + e column number =
     (0, 1, 2, 5)
  coefficient of determination =  0.7712123206719487
```

```
  durbin watson =   1.2458334878080959
  multicollinearity =    971.6342737268205
  heteroscedasticity =   (10962.878090807526, 0.0,
3420.8936295660956, 0.0)
  model: y = b0 + b1x1A2 + b2x2A2 + ... + e column number =     (0,
1, 2, 6)
  coefficient of determination =   0.7715272297685392
  durbin watson =   1.2471376459558166
  multicollinearity =    83.15048207452044
  heteroscedasticity =   (11047.349312096749, 0.0,
3453.8592930875507, 0.0)
  model: y = b0 + b1x1A2 + b2x2A2 + ... + e column number =     (0,
1, 2, 7)
  coefficient of determination =   0.7711835814261333
  durbin watson =   1.2457476421434832
  multicollinearity =    669.1623620207082
  heteroscedasticity =   (10988.705359642558, 0.0,
3430.959562850149, 0.0)
  model: y = b0 + b1x1A2 + b2x2A2 + ... + e column number =     (0,
1, 2, 8)
  coefficient of determination =   0.7712252904342234
  durbin watson =   1.244665100126957
  multicollinearity =    26872.136761657843
```